**B3**

**PCT**

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

| | | |
|---|---|---|
| (51) International Patent Classification [7] : C12Q 1/68 | A2 | (11) International Publication Number: **WO 00/60124** |
| | | (43) International Publication Date: 12 October 2000 (12.10.00) |

(21) International Application Number: PCT/US00/09284

(22) International Filing Date: 6 April 2000 (06.04.00)

(30) Priority Data:
60/127,932     6 April 1999 (06.04.99)     US

(71) Applicant: YALE UNIVERSITY [US/US]; 451 College Street, New Haven, CT 06511 (US).

(72) Inventors: LIZARDI, Paul, M.; 7 Stoney Brook Road, Wallingford, CT 06492 (US). ROTH, Matthew, E.; 20 Medley Lane, Branford, CT 06405 (US). FENG, Li; 1579 Rhinelander Avenue, Apartment 2T, Bronx, NY 10461 (US). GUERRA, Cesar, E.; 441 Nortontown Road, Guilford, CT 06437 (US). WEBER, Shane, C.; 1 Evergreen Drive, Woodbridge, CT 06525 (US). KAUFMAN, Joseph, C.; 24 Farmington Road, Hamden, CT 06518 (US). LATIMER, Darin, R.; 123 Florence Road, Apartment 2B, Branford, CT 06405 (US).

(74) Agents: PABST, Patrea, L. et al.; Arnall Golden & Gregory, LLP, 2800 One Atlantic Center, 1201 West Peachtree Street, Atlanta, GA 30309–3450 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published**
*Without international search report and to be republished upon receipt of that report.*

---

(54) Title: FIXED ADDRESS ANALYSIS OF SEQUENCE TAGS

(57) Abstract

Disclosed is a method for the comprehensive analysis of nucleic acid samples and a detector composition for use in the method. The method, referred to as Fixed Address Analysis of Sequence Tags (FAAST), involves generation of a set of nucleic acid fragments having a variety of sticky end sequences; indexing of the fragments into sets based on the sequence of sticky ends; associating a detector sequence with the fragments; sequence–based capture of the indexed fragments on a detector array; and detection of the fragment labels. Generation of the multiple sticky end sequences is accomplished by incubating the nucleic acid sample with one or more nucleic acid cleaving reagents. The indexed fragments are captured by hybridization and coupling, preferably by ligation, to a probe. The method allows a complex sample of nucleic acid to be quickly and easily cataloged in a reproducible and sequence–specific manner. One form of the method allows determination of associations, in a nucleic acid molecule, of different combinations of known or potential sequences. Another form of the method assesses modification of sequences in nucleic acid molecules by basing cleavage of the molecules on the presence or absence of modification.

# FIXED ADDRESS ANALYSIS OF SEQUENCE TAGS

## BACKGROUND OF THE INVENTION

The disclosed invention is generally in the field of nucleic acid characterization and analysis, and specifically in the area of analysis and comparison of gene expression patterns and genomes.

The study of differences in gene-expression patterns is one of the most promising approaches for understanding mechanisms of differentiation and development. In addition, the identification of disease-related target molecules opens new avenues for rational pharmaceutical intervention. Currently, there are two main approaches to the analysis of molecular expression patterns: (1) the generation of mRNA-expression maps and (2) examination of the 'proteome', in which the expression profile of proteins is analyzed by techniques such as two-dimensional gel electrophoresis, mass spectrometry [matrix-assisted-desorption-ionization-time-of-flight (MALDI-TOF) or electrospray] and by the ability to sequence sub-picomole amounts of protein. Classical approaches to transcript imaging, such as northern blotting or plaque hybridization, are time-consuming and material-intensive ways to analyze mRNA-expression patterns. For these reasons, other methods for high-throughput screening in industrial and clinical research have been developed.

A breakthrough in the analysis of gene expression was the development of the northern-blot technique in 1977 (Alwine et al., *Proc. Natl. Acad. Sci. U.S.A.* 74:5350-5354 (1977)). With this technique, labeled cDNA or RNA probes are hybridized to RNA blots to study the expression patterns of mRNA transcripts. Alternatively, RNase-protection assays can detect the expression of specific RNAs. These assays allow the expression of mRNA subsets to be determined in a parallel manner. For RNase-protection assays, the sequence of the analyzed mRNA has to be known in order to synthesize a labeled cDNA that forms a hybrid with the selected mRNA; such hybrids resist RNA degradation by a single-strand-specific nuclease and can be detected by gel electrophoresis. As a third approach, differential plaque-filter hybridization allows the identification of specific differences in the expression of cloned cDNAs (Maniatis *et al. Cell* 15:687-701 (1978)). Although all of these techniques are

1

excellent tools for studying differences in gene expression, the limiting factor of these classical methods is that expression patterns can be analyzed only for known genes.

5      The analysis of gene-expression patterns made a significant advance with the development of subtractive cDNA libraries, which are generated by hybridizing an mRNA pool of one origin to an mRNA pool of a different origin. Transcripts that do not find a complementary strand in the hybridization step are then used for the construction of a cDNA library (Hedrick et al., *Nature* 308:149-153 (1984)). A variety of refinements to this method have been

10     developed to identify specific mRNAs (Swaroop et al., *Nucleic Acids Res.* 25:1954 (1991); Diatchenko *et al*, *Proc. Natl. Acad. Sci. U.S.A* 93:6025-6030 (1996)). One of these is the selective amplification of differentially expressed mRNAs via biotin- and restriction-mediated enrichment (SABRE; Lavery et al., *Proc. Natl. Acad. Sci. U.S.A.* 94:6831-6836 (1997)), cDNAs derived from a

15     tester population are hybridized against the cDNAs of a driver (control) population. After a purification step specific for tester-cDNA-containing hybrids, tester-tester homohybrids are specifically amplified using an added linker, thus allowing the isolation of previously unknown genes.

The technique of differential display of eukaryotic mRNA was the first

20     one-tube method to analyze and compare transcribed genes systematically in a bi-directional fashion; subtractive and differential hybridization techniques have only been adapted for the unidirectional identification of differentially expressed genes (Liang and Pardee, *Science* 257:967-971 (1992)). Refinements have been proposed to strengthen reproducibility, efficiency, and performance

25     of differential display (Bauer *et al.*, *Nucleic Acids Res.* 11:4272-4280 (1993); Liang and Pardee, *Curr. Opin. Immunol* 7:274-280 (1995); Ito and Sakaki, *Methods Mol. Biol.* 85:37-44 (1997); Praschar and Weissman, *Proc. Natl. Acad. Sci U.S.A.* 93;659-663 (1996)). Although these approaches are more reproducible and precise than traditional PCR-based differential display, they

30     still require the use of gel electrophoresis, and often implies the exclusion of certain DNA fragments from analysis.

Originally developed to identify differences between two complex genomes, representational difference analysis (RDA) was adapted to analyze differential gene expression by taking advantage of both subtractive hybridization and PCR (Lisitsyn et al., *Science* 259:946-951 (1993); Hubank

5 and Schatz, *Nucleic Acids Res.* 22:5640-5648 (1994)). In the first step, mRNA derived from two different populations, the tester and the driver (control), is reverse transcribed; the tester cDNA represents the cDNA population in which differential gene expression is expected to occur. Following digestion with a frequently cutting restriction endonuclease, linkers are ligated to both ends of

10 the cDNA. A PCR step then generates the initial representation of the different gene pools. The linkers of the tester and driver cDNA are digested and a new linker is ligated to the ends of the tester cDNA. The tester and driver cDNAs are then mixed in a 1:100 ratio with an excess of driver cDNA in order to promote hybridization between single-stranded cDNAs common in both tester

15 and driver cDNA pools. Following hybridization of the cDNAs, a PCR exponentially amplifies only those homoduplexes generated by the tester cDNA, via the priming sites on both ends of the double-stranded cDNA (O'Neill and Sinclair, *Nucleic Acids Res.* 25:2681-2682 (1997); Wada et al., *Kidney Int.* 51:1629-1638 (1997); Edman et al., *J.* 323:113-118 (1997)).

20        The gene-expression pattern of a cell or organism determines its basic biological characteristics. In order to accelerate the discovery and characterization of mRNA-encoding sequences, the idea emerged to sequence fragments of cDNA randomly, direct from a variety of tissues (Adams *et al.*, *Science* 252:1651-1656 (1991); Adams *et al.*, *Nature* 377:3-16 (1995)). These

25 expressed sequence tags (ESTs) allow the identification of coding regions in genome-derived sequences. Publicly available EST databases allow the comparative analysis of gene expression by computer. Differentially expressed genes can be identified by comparing the databases of expressed sequence tags of a given organ or cell type with sequence information from a different origin

30 (Lee *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 92:8303-8307 (1995); Vasmatzis et al., *Proc. Natl. Acad. Sci. U. S. A.* 95:300-304 (1998)). A drawback to sequencing of ESTs is the requirement for large-scale sequencing facilities.

3

Serial analysis of gene expression (SAGE) is a sequence-based approach to the identification of differentially expressed genes through comparative analyses (Velculescu et al., *Science* 270:484-487 (1995)). It allows the simultaneous analysis of sequences that derive from different cell population or

5      tissues. Three steps form the molecular basis for SAGE: (1) generation of a sequence tag (10-14 bp) to identify expressed transcripts; (2) ligation of sequence tags to obtain concatemers that can be cloned and sequenced; and (3) comparison of the sequence data to determine differences in expression of genes that have been identified by the tags. This procedure is performed for every

10     mRNA population to be analyzed. A major drawback of SAGE is the fact that corresponding genes can be identified only for those tags that are deposited in gene banks, thus making the efficiency of SAGE dependent on the extent of available databases. Alternatively, a major sequencing effort is required to complete a SAGE data set capable of providing 95 % coverage of any given

15     mRNA population, simply because most of the sequencing work yields repetitive reads on those tags that are present in high frequency in cellular mRNA. In other words, SAGE sequencing experiments yield diminishing returns for rare mRNAs, whose unique tags will begin to accumulate in the database only after many weeks of sequencing effort.

20            A different approach to the study of gene-expression profiles and genome composition is the use of DNA microarrays. Current DNA microarrays are systematically gridded at high density. Such microarrays are generated by using cDNAs (for example, ESTs), PCR products or cloned DNA, which are linked to the surface of nylon filters, glass slides or silicon chips (Schena et al.,

25     *Science* 270, 467-470 (1995)). DNA arrays can also be assembled from synthetic oligonucleotides, either by directly applying the synthesized oligonucleotides, either by directly applying the synthesized oligonucleotides to the matrix or by a more sophisticated method that combines photolithography and solid-phase chemical synthesis (Fodor et al., *Nature* 364:555-556 (1993)).

30     To determine differences in gene-expression, labeled cDNAs or oligonucleotides are hybridized to the DNA- or oligomer-carrying arrays.

4

When using different fluorophores for labeling cDNAs or oligonucleotides, two probes can be applied simultaneously to the array and compared at different wavelengths. The expression of 10,000 genes and more can be analyzed on a single chip (Chee *et al.*, *Science* 274:610-614 (1996)). However, depending on

5      the sensitivity of both cDNA and oligonucleotide arrays, the intensity of hybridization signals can leave the linear range when either weakly or abundantly expressed genes are analyzed. Thus, individual optimization steps are required to ensure the accurate detection of differentially expressed genes. While such microarray methods may be used to address a number of interesting

10     biological questions, they are not suitable for the discovery of new genes.

There is a need for a method that combines the power and convenience of array hybridization technology with the capability for gene discovery inherent in differential display or SAGE. Such a method would be most attractive if it could enable comprehensive gene expression analysis without the

15     use of gel electrophoresis, and without the need for a redundant DNA sequencing effort.

Therefore, it is an object of the present invention to provide a method for the comprehensive analysis of nucleic acid sequence tags.

It is another object of the present invention to provide a detector

20     composition that allows indexing of nucleic acid sequence tags.

## BRIEF SUMMARY OF THE INVENTION

Disclosed is a method for the comprehensive analysis of nucleic acid samples and a detector composition for use in the method. The method, referred to as Fixed Address Analysis of Sequence Tags (FAAST), involves

25     generation of a set of nucleic acid fragments having a variety of sticky end sequences; indexing of the fragments into sets based on the sequence of sticky ends; associating a detector sequence with the fragments; sequence-based capture of the indexed fragments on a detector array; and detection of the fragment labels. Generation of the multiple sticky end sequences is

30     accomplished by incubating the nucleic acid sample with one or more nucleic acid cleaving reagents. Preferably this is accomplished by subjecting the nucleic acid sample to digestion by a restriction endonuclease that cleaves at a

5

site different from the recognition sequence, or by multiple restriction endonucleases. The indexed fragments are captured by hybridization and coupling, preferably by ligation, to a probe. The probe is preferably immobilized in an array or on sortable beads.

5          The method allows detection of the indexed fragments where detection provides some sequence information for the fragments including the sequence of the original sticky end of each fragment, the recognition sequence of the restriction endonuclease (if different from the sticky end sequence), and the sequence corresponding to the probe. The method allows a complex sample of

10        nucleic acid to be cataloged quickly and easily in a reproducible and sequence-specific manner.

One form of the FAAST method, referred to as variable address analysis of sequence tags (VAAST) allows determination of associations, in a nucleic acid molecule, of different combinations of known or potential sequences. For

15        example, particular combinations of joining and variable regions in immunoglobulins or T cell receptors can be determined. Another form of the FAAST method, referred to as modification assisted analysis of sequence tags (MAAST), assesses modification of sequences in nucleic acid molecules by basing cleavage of the molecules on the presence or absence of modification.

20        For example, a site that is methylated in a nucleic acid molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is insensitive to methylation will cleave at that site, thus producing a different pattern of sequence tags.

## BRIEF DESCRIPTION OF THE DRAWINGS

25        Figure 1 is a listing of examples of ligator-detectors (numbered sequences) designed for use with one of two example adaptor-indexers (top). The sticky end sequences (or their complements) are shown in bold.

Figure 2 is a list of nucleic acid molecules used and formed during an example of the disclosed method focused on a single example nucleic acid

30        molecule (panel A). The restriction enzyme recognition sequence is underlined and the sticky end sequence is in bold. Panel B shows the nucleic acid molecule after cleavage with FokI. Panel C shows the nucleic acid fragment (left) and an

6

example of a compatible adaptor-indexer (right; this is adaptor-indexer I from Figure 1). Panel D shows the adaptor-indexer ligated to the nucleic acid fragment. Panel E shows one strand (the top strand) of the adaptor/fragment (top) and an example of a compatible ligator-detector (bottom; this is ligator-

5      detector 13 from Figure 1). Panel F shows the ligator-detector hybridized to the adaptor/fragment strand. Panel G shows the adaptor/fragment/ligator (top) and a compatible hexamer detector array probe (bottom) immobilized on a detector array. Panel H shows the detector array probe ligated to the ligator-detector (both hybridized to the adaptor/fragment). Panel I show the ligated

10     probe/ligator-detector, which is immobilized on the detector array.

Figure 3 is a list of nucleic acid molecules used and formed during an example of the disclosed method focused on a single example nucleic acid molecule (panel A). In this example, the bottom strand of the nucleic acid molecule is used. The restriction enzyme recognition sequence is underlined

15     and the sticky end sequence is in bold. Panel B shows the nucleic acid molecule after cleavage with FokI. Panel C shows the nucleic acid fragment (left) and an example of a compatible adaptor-indexer (right; this is adaptor-indexer I from Figure 1). Panel D shows the adaptor-indexer ligated to the nucleic acid fragment. Panel E shows the bottom strand of the adaptor/fragment (top) and an

20     example of a compatible ligator-detector (bottom; this is ligator-detector 1 from Figure 1). Panel F shows the ligator-detector hybridized to the adaptor/fragment strand. Panel G shows the adaptor/fragment/ligator (top) and a compatible hexamer detector array probe (bottom) immobilized on a detector array. Panel H shows the detector array probe ligated to the ligator-detector

25     (both hybridized to the adaptor/fragment). Panel I show the ligated probe/ligator-detector, which is immobilized on the detector array.

Figure 4 is a list of nucleic acid molecules used and formed during an example of the disclosed method focused on a single example nucleic acid molecule (panel A). This is the same as the example in Figure 2 except the

30     addition of a second adaptor is illustrated. The restriction enzyme recognition sequence is underlined and the sticky end sequences are in bold. Panel D shows the adaptor-indexer ligated to the nucleic acid fragment. Panel D1 shows the

7

nucleic acid molecule after cleavage with Sau3AI. Panel D2 shows the nucleic acid fragment (right) and an example of a compatible second adaptor (left). Panel D3 shows the second adaptor ligated to the nucleic acid fragment (the adaptor-indexer is ligated to the other end of the fragment). Panel F shows a

5      ligator-detector hybridized to a strand of the adaptor/fragment. Panel H shows a detector array probe ligated to the ligator-detector (both hybridized to the adaptor/fragment). Panel I show the ligated probe/ligator-detector, which is immobilized on the detector array.

Figure 5 is a list of nucleic acid molecules used and formed during an

10     example of the disclosed method focused on a single example nucleic acid molecule (panel A). This is the same as the example in Figure 2 except the addition of a capture tag is illustrated. The restriction enzyme recognition sequence is underlined and the sticky end sequence is in bold. Panel B shows the nucleic acid molecule after cleavage with FokI. Panel C shows the nucleic

15     acid fragment (left) and an example of a compatible adaptor-indexer (right; this is adaptor-indexer 1 from Figure 1). The adaptor-indexer includes a capture tag (tag). Panel D shows the adaptor-indexer ligated to the nucleic acid fragment. Panel E shows one strand (the top strand) of the adaptor/fragment (top) and an example of a compatible ligator-detector (bottom; this is ligator-detector 13

20     from Figure 1). Separation of the strands can be facilitated by the capture tag. Panel F shows the ligator-detector hybridized to the adaptor/fragment strand. Panel G shows the adaptor/fragment/ligator (top) and a compatible hexamer detector array probe (bottom) immobilized on a detector array. Panel H shows the detector array probe ligated to the ligator-detector (both hybridized to the

25     adaptor/fragment). Panel I show the ligated probe/ligator-detector, which is immobilized on the detector array.

## DETAILED DESCRIPTION OF THE INVENTION

The disclosed method, referred to as Fixed Address Analysis of Sequence Tags (FAAST), allows a complex sample of nucleic acid to be

30     quickly and easily cataloged in a reproducible and sequence-specific manner. Such a catalog can be compared with other, similarly prepared catalogs of other nucleic acid samples to allow convenient detection of differences between the

samples.  The catalogs, which incorporate a significant amount of information about the nucleic acid samples, can serve as fingerprints of the nucleic acid samples which can be used both for detection of related nucleic acid samples and comparison of nucleic acid samples.  For example, the presence or identity

5      of specific organisms can be detected by producing a catalog of nucleic acid of the test organism and comparing the resulting catalog with reference catalogs prepared from known organisms.  Changes and differences in gene expression patterns can also be detected by preparing catalogs of mRNA from different cell samples and comparing the catalogs.  The catalog of sequences can also be used

10     to produce a set of probes or primers that is specific for the source of a nucleic acid sample.

Comparison of nucleic acid catalogs produced with the disclosed method is facilitated by the highly ordered nature of the sequence information produced and cataloged in the method.  Use of immobilization, sorting, and/or

15     array detection in the method allows automation of the method, the cataloging of the information, and comparisons to other catalogs.  The method results in the equivalent of a large number of sequence-specific bins that can be filled, empty, or filled to different levels, with the pattern of filled and empty bins, and/or of the amount of signal in a bin, providing information about the nucleic

20     acid sample that has been cataloged.

The FAAST method involves the following basic steps.  A nucleic acid sample is incubated with one or more nucleic acid cleaving reagents, preferably restriction endonucleases, that results in a set of DNA fragments having sticky ends with a variety of sequences.  In a preferred form of the method, the sample

25     can be divided into aliquots (referred to as index samples); preferably as many aliquots as there are sticky end sequences.  Where multiple restriction endonucleases are used, the nucleic acid sample is preferably divided into index samples before cleavage.  Where a single restriction endonuclease is used, the nucleic acid sample is preferably divided into index samples following

30     cleavage.  Each sample (each index sample if the nucleic acid sample was divided) is then mixed with one or more adaptor-indexers, each of which has a sticky end compatible with one of the possible sticky ends on the DNA

9

fragments in that index sample. It is preferred that a different adaptor-indexer be mixed with each index sample. The adaptor-indexes are then covalently coupled, preferably by ligation, onto compatible DNA fragments.

5      The DNA fragments in each sample (or each index sample) are then hybridized to ligator-detectors. One end of each ligator-detector has sequence matching or complementary to all or part of one of the possible sticky end sequences generated by the first nucleic acid cleaving reagent(s), or all or part of the recognition sequence of the first nucleic acid cleaving reagent(s) when cleavage is not offset from the recognition sequence. The ligator-detector can,

10     and preferably does, have sequence matching or complementary to sequence adjacent to the sticky end sequence in the fragment coupled to the adaptor-indexer. The ligator-detector used in each index sample preferably matches or is complementary to sequence, including sticky end sequence, in the adaptor-indexer sequence used in that index sample.

15     Finally, each sample (or index sample) is reacted with and coupled, preferably by ligation, to one or more detector probes. Preferably, the set of detector probes used include every possible sequence of a given length (for example, every possible six base sequence). The ends of the probes and the ligator-detector oligonucleotides are coupled only if the probe hybridizes

20     adjacent to the end of the ligator-adaptor. The probes are preferably immobilized oligonucleotides.

Each processed DNA fragment from the sample will result in a signal based on coupling of the ligator-detector to a probe. A complex nucleic acid sample will produce a unique pattern of signals. It is this pattern that allows

25     unique cataloging of nucleic acid samples and sensitive and powerful comparisons of the patterns of signals produced from different nucleic acid samples.

Coupling of ligator-detectors to probes can be detected directly or indirectly. For example, any of the probe, the ligator-detector, or the associated

30     adaptor-indexer can be detected. Association of a ligator-detector or adaptor-indexer with a given probe is indicative of coupling of the probe and ligator-detector. Detection of such associations can be facilitated through

10

immobilization of the probes, detector-ligators, or adaptor-indexers, and

through the use of capture tags, sorting tags and detectable labels in association

with the probes, detector-ligators, and/or adaptor-indexers. Any combination of

immobilization and association with capture tags, sorting tags, and labels can be

5      used. Preferably, the probes are immobilized in arrays and the ligator-detectors

are associated with a detectable label. Thus, detection of a signal at a particular

location in a particular array of immobilized probes can provide information

about nucleic acid fragments indexed from the nucleic acid sample.

Where the probes are immobilized in arrays, the array, and location in

10     the array, where a DNA fragment generates a signal identify the sequence of the

sticky end of the DNA fragment and of the sequence adjacent to the sticky end

(or the recognition sequence of the first restriction enzyme and adjacent

sequence when the recognition sequence and sticky end sequence overlap).

This is a ten base sequence when a four base sticky end and six base

15     immobilized oligonucleotides are used. The fixed relationship between the

recognition sequence and the cleavage site of the type II-S restriction enzyme,

when used, and the identity of the recognition sequence, provide additional

sequence information about the DNA fragment. The same effect can be

accomplished by otherwise capturing, sorting, or detecting particular probes

20     (via capture tags, sorting tags, and labels). That is, so long as the probe and the

ligator-detector coupled to it can be identified, a pattern can be determined.

Optionally, prior to hybridization with ligator-detectors, each sample (or

index sample) can be incubated with one or more other nucleic acid cleaving

reagents (referred to as second nucleic acid cleaving reagents), preferably a

25     restriction enzyme having a four base recognition sequence. A second adaptor

can then be coupled, preferably by ligation, to the DNA fragments in the

samples. The DNA fragments can then be amplified using any suitable method,

such as PCR.

One form of the FAAST method, referred to as modification assisted

30     analysis of sequence tags (MAAST), assesses modification of sequences in

nucleic acid molecules by basing cleavage of the molecules on the presence or

absence of modification. For example, a site that is methylated in a nucleic acid

11

molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is insensitive to methylation will cleave at that site, thus producing a different pattern of sequence tags.

Another form of the FAAST method, referred to as variable address analysis of sequence tags (VAAST) allows determination of associations, in a nucleic acid molecule, of different combinations of known or potential sequences. For example, particular combinations of joining and variable regions in immunoglobulins or T cell receptors can be determined. VAAST uses the same basic steps as FAAST and adds a step prior to cleavage of the nucleic acid sample. In VAAST, a recognition site for cleavage, preferably a restriction enzyme recognition site, is introduced into nucleic acid fragments in the nucleic acid sample. This recognition site is then used as the target of cleavage in the first basic FAAST method. The adaptor-indexers should be chosen to match known or potential sequences that would appear adjacent to the sequence into which the recognition site was introduced. The result is fragments with defined end sequences surrounding a central sequence derived from a nucleic acid fragment. This allows the association of known or potential sequences to be assessed. In particular, the association of the sequence into which the recognition site was introduced with a particular adaptor-indexer (which has sequence matching the known or potential adjacent sequence) can be detected.

## Materials

**Nucleic Acid Samples**

Any nucleic acid sample can be used with the disclosed method. Examples of suitable nucleic acid samples include genomic samples, mRNA samples, cDNA samples, nucleic acid libraries (including cDNA and genomic libraries), whole cell samples, environmental samples, culture samples, tissue samples, bodily fluids, and biopsy samples. Numerous other sources of nucleic acid samples are known or can be developed and any can be used with the disclosed method. Preferred nucleic acid samples for use with the disclosed method are nucleic acid samples of significant complexity such as genomic samples and mRNA samples.

12

Nucleic acid fragments are segments of larger nucleic molecules. Nucleic acid fragments, as used in the disclosed method, generally refer to nucleic acid molecules that have been cleaved. A nucleic acid sample that has been incubated with a nucleic acid cleaving reagent is referred to as a digested

5     sample. A nucleic acid sample that has been digested using a restriction enzyme is referred to as a digested sample.

An index sample is a nucleic acid sample that has been divided into different aliquots for further processing. In the context of the disclosed method, index samples are preferably aliquots of a digested nucleic acid sample to which

10     different adaptor-indexers are added for coupling, preferably by ligation, to nucleic acid fragments present in the digested sample. In the disclosed method, different nucleic acid fragments are processed in the different index samples based on the sticky end sequence of the fragments. Thus, it is preferred that digested nucleic acid samples be divided into as many index samples as the

15     number of possible sticky end sequences generated by the nucleic acid cleaving reagent used to digest the sample. Where multiple different nucleic acid cleaving reagents are used to cleave a nucleic acid sample, it is preferred that the nucleic acid sample be divided into as many aliquots as nucleic acid cleaving reagents used and that the nucleic acid sample be divided prior to

20     cleavage.

A control nucleic acid sample is a nucleic acid sample to which another nucleic acid sample (which can be referred to as a tester nucleic acid sample) is to be compared. A control index sample is an index sample to which another index sample (which can be referred to as a tester index sample) is to be

25     compared.

Secondary index samples are aliquots of index samples. Thus, index samples can be divided into a plurality of secondary index samples. Secondary index samples are to be cleaved with a nucleic acid cleaving reagent, preferably a restriction enzyme. Restricted index samples and non-restricted index

30     samples are aliquots of index samples. Restricted index samples are to be cleaved with a nucleic acid cleaving reagent while non-restricted index samples are not. Restricted secondary index samples and non-restricted secondary index

13

samples are aliquots of secondary index samples. Restricted secondary index samples are to be cleaved with a nucleic acid cleaving reagent while non-restricted secondary index samples are not. Secondary index samples, restricted index samples, non-restricted index samples, restricted secondary index

5      samples, and non-restricted secondary index samples are referred to collectively herein as derivative index samples. Each is derived from an index sample and, in some cases, from another derivative index sample.

**Nucleic Acid Cleaving Reagents**

The disclosed method makes use of nucleic acid cleaving reagents.

10     Nucleic acid cleaving reagents are compounds, complexes, and enzymes that cause, mediate, or catalyze cleavage on nucleic acid molecules. Preferred nucleic acid cleaving reagents are those that cleave nucleic acid molecules in a sequence-specific manner. Restriction enzymes (also referred to as restriction endonucleases) are the preferred form of nucleic acid cleaving reagents. Other

15     nucleic acid cleaving reagents include the universal restriction endonucleases of Szybalski (Szybalski, Gene 40(2-3):169-73 (1985); Podhajska and Szybalski, Gene 40(2-3):175-82 (1985)[published erratum appears in Gene 43(3):325 (1985)]), and the advanced DNA cleavage systems have been evolved by Breaker et al. (Carmi et al., Proc Natl Acad Sci U S A 95(5):2233-2237 (1998)).

20     Many nucleic acid cleaving reagents are known and can be used with the disclosed method. Relevant to the disclosed method, nucleic acid cleaving reagents generally have a recognition sequence and a cleavage site. Many nucleic acid cleaving reagents, especially restriction enzymes, also generate sticky ends at the cleavage site. A recognition sequence is the nucleotide

25     sequence which, if present in a nucleic acid molecule, will direct cleavage of the nucleic acid molecule by a cognate nucleic acid cleaving reagent. The cleavage site of a nucleic acid cleaving reagent is the site, usually in relation to the recognition sequence, where the nucleic acid cleaving reagent cleaves a nucleic acid molecule. Sticky ends (also referred to as cohesive ends, protruding ends,

30     and 5' or 3' overhangs) are single-stranded nucleic acid segments at the end of a double-stranded nucleic acid segment.

14

For specific embodiments of the method, the nucleic acid cleaving reagents used will have certain properties and/or certain relationships to other restriction enzymes used in the method. For example, in some preferred embodiments of the disclosed method, nucleic acid cleaving reagents that

5      generates sticky ends having a plurality of different sequences are preferred, with nucleic acid cleaving reagents having a cleavage site offset from the recognition sequence being most preferred. Other embodiments of the disclosed method require the use of different nucleic acid cleaving reagents that have different recognition sequences and/or generate different sticky ends than

10     other nucleic acid cleaving reagents used on the same index sample at other stages in the method. For example, where three digests (that is, cleavage reactions) are used in the method, it is preferred that the nucleic acid cleaving reagents used in each of the digests have a recognition sequence different from that of the nucleic acid cleaving reagents used in the other digests. In such

15     cases, the known properties of nucleic acid cleaving reagents can be used to select or design appropriate nucleic acid cleaving reagents.

Where a nucleic acid cleaving reagent cleaves DNA at a site different or offset from the recognition sequence, a variety of sticky ends having different sequences can be generated. This is because recognition sequences in nucleic

20     acids can occur next to any sequence and therefore the site of cleavage can have any sequence. For example, FokI cleaves 9 (upper strand) and 13 (lower strand) nucleotides downstream from the recognition site of GGATG. The four base sticky end will have whatever sequence happens to be 10 to 13 nucleotides away from the recognition site. Given enough cleavage sites, a total of 256

25     different sticky end sequences (that is every possible four base sequence) can result from a FokI digestion. As a result, restriction enzymes such as type II-S restriction enzymes can be said to generate sticky ends having a plurality of different sequences.

As used herein, unless otherwise indicated, the terms digest, digestion,

30     digested, and digesting refer generally to a cleavage reaction or the act of cleaving and is not intended to be limited to cleavage by a protein enzyme or by any particular mechanism. Similarly, the term restricted is intended to refer to

15

any nucleic acid cleavage, not just cleavage by a restriction enzyme. In the context of nucleic acid cleaving reagents, sequence-specific requires only some sequence specificity, not absolute sequence specificity. That is, nucleic acid cleaving reagents having a completely or partially defined recognition sequence

5      are preferred. Thus, nucleic acid cleaving reagents having some degeneracy in their recognition sequence are still considered sequence-specific.

A second nucleic acid cleaving reagent is a nucleic acid cleaving reagent used to digest a secondary index sample. A third nucleic acid cleaving reagent is an nucleic acid cleaving reagent used to digest a restricted index sample or a

10     restricted secondary index sample. Second and third nucleic acid cleaving reagents are preferably type II restriction endonucleases that cleave in the recognition sequence. A second restriction enzyme is a restriction enzyme used to digest a secondary index sample. A third restriction enzyme is an enzyme used to digest a restricted index sample or a restricted secondary index sample.

15     Second and third restriction enzymes are preferably type II restriction endonucleases that cleave in the recognition sequence.

In addition to the use of restriction enzymes in a standard mode, one can make use of the type II-S enzymes as universal restriction endonuclease as described by Szybalski (Szybalski, Gene 40(2-3):169-73 (1985); Podhajska and

20     Szybalski, Gene 40(2-3):175-82 (1985)[published erratum appears in Gene 43(3):325 (1985)]). In the Szybalski technique single stranded or double stranded DNA can be cleaved at any arbitrary (but specific) site utilizing the structure described in combination with a type II-S enzyme. More advanced DNA cleavage systems have been evolved by Breaker et al. (Carmi et al., Proc

25     Natl Acad Sci U S A 95(5):2233-2237 (1998)). In these systems Breaker has shown that DNA recognize a particular sequence in a target DNA and can cleave the target DNA, single stranded or double stranded targets. With Breaker's system for evolution of DNA for a particular action, it is clear that given reasonable time and effort a suitable DNA for a recognition and particular

30     cleavage result is practical.

16

### Adaptor-Indexers

Adaptor-indexers are double-stranded nucleic acids containing a single-stranded portion and a double-stranded portion. The single-stranded portion is at one end of the adaptor-indexer and constitutes a sticky end. The sticky end is

5      referred to as the sticky end portion of the adaptor-indexer. It is preferable that the protruding single strand (sticky end) have two, three, four, or five nucleotides. The double-stranded portion of adaptor-indexers may have any convenient sequence or length. In general, the sequence and length of the double-stranded portion is selected to be adapted to subsequent steps in the

10     method. For example, sequences in the adaptor-indexer may be used for primer or probe hybridization. A main purpose of adaptor-indexers is to provide sequence for hybridization by a ligator-detector. If the fragments in the samples to which adaptors have been coupled are to be amplified, the adaptor-indexers can also provide sequence for primer hybridization (which can overlap or be

15     contiguous with sequence for ligator-detector hybridization). Thus, preferred sequence composition and length for the double-stranded portion of adaptor-indexers will generally be those that are useful for probe and primer hybridization. Adaptor-indexers can also include a detector portion which is designed to facilitate detection of the adaptor-indexer. The detection portion

20     can be, for example, a sequence that is a hybridization target or it can be a label or tag.

Generally, the sequence of the double-stranded portion of an adaptor-indexer should not include the recognition sequence of any restriction enzyme to be used in a subsequent step in the method. It is preferred that adaptor-

25     indexers not have any sequences that are self-complementary. It is considered that this condition is met if there are no complementary regions greater than six nucleotides long without a mismatch or gap.

A set of adaptor-indexers for use in the disclosed method should include different adaptor-indexers where the single-stranded portion each have a

30     different nucleotide sequence selected from combinations and permutations of the nucleotides A, C, G, and T. Where multiple nucleic acid cleaving reagents are used in the first digest, the single-stranded portion of each adaptor-indexer

17

can have a different nucleotide sequence compatible with a sticky end sequence generated by one of the nucleic acid cleaving reagents. While the sticky ends of adaptor-indexers in one set have different sequences, it is preferred that they be of the same length to facilitate use of the set to index fragments produced by

5      cleavage by one nucleic acid cleaving reagent. It is preferable that the members of a set of adaptor-indexers contain a double-stranded portion which is identical for each member of the set.

A preferred set of indexing linker strands comprising: (a) at least two single-stranded first oligonucleotides each having a common identical sequence,

10    and a unique sequence of a length selected from 2, 3, 4 and 5 nucleotides selected from permutations and combinations of A, G, C and T nucleotides, at one end selected from a 3' end and a 5' end; and (b) a single stranded second oligonucleotide whose sequence is complementary to the common sequence of the first oligonucleotides such that, when hybridized with any one of the first

15    oligonucleotides, a double-stranded adaptor-indexer would result which includes an end having a sticky end with a unique sequence.

Adaptor-indexers can also contain or be associated with capture tags to facilitate immobilization or capture of fragments to which adaptor-indexers have been coupled. In general, the capture tag can be one member of a binding

20    pair such as biotin and streptavidin. Capture tags are discussed more fully elsewhere herein. Adaptor-indexers can also contain or be associated with sorting tags to facilitate sorting or separation of fragments to which adaptor-indexers have been coupled. In general, the sorting tag can be a detectable label such as a fluorescent moiety or a manipulable moiety such as a magnetic bead.

25    Sorting tags are discussed more fully elsewhere herein. Adaptor-indexers can also contain or be associated with labels to facilitate detection of fragments to which adaptor-indexers have been coupled. Adaptor-indexers can also be immobilized on a substrate.

Adaptor-indexers can also include a protruding end at the end opposite

30    the sticky end. Such an end can be used as, for example, a hybridization target for a label to be associated with the adaptor-indexer (and thus can be considered the detection portion of the adaptor-indexer). Adaptor-indexers can also include

18

one or more photocleavable nucleotides to facilitate release of adaptor-indexer sequences for detection. Photocleavable nucleotides and their use are described in WO 00/04036.

Adaptor-indexers need not be composed of naturally occurring
5      nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the adaptor-indexer have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

**Ligator-Detectors**

10      A ligator-detector is a nucleic acid molecule containing a single-stranded region that is complementary to a portion of a nucleic acid fragment generated in the disclosed method from a nucleic acid sample. The ligator-detectors generally have a specific sequence relationship to adaptor-indexers. Ligator-detectors include sequence--referred to as the detector portion of the
15      ligator-detector--matching or complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers. Thus, the ligator-detector can have sequence matching or complementary to sequence in the nucleic acid fragment adjacent to the sticky end sequence (on either or both sides), matching or complementary to the sticky end, or sequence
20      matching or complementary to both sequence in the nucleic acid fragment adjacent to the sticky end sequence and the sticky end.

Preferably, the sequence of a ligator-detector matches or is complementary to all or part of a sticky end sequence and all or part of the adjacent sequence of the adaptor-indexer designed for use with that sticky end
25      sequence. In this form, the sequence of a ligator-detector matches or is complementary to all or part of the recognition sequence of the first restriction enzyme(s) when cleavage is not offset from the recognition sequence. For VAAST, it is preferred that the ligator-detector include sequence matching or complementary to sequence in the nucleic acid fragment adjacent to the sticky
30      end sequence. Whether the sequence in the ligator-detector is matching or complementary determines which strand of the adaptor-indexer and/or fragment will hybridize to the detector-ligator. It is preferred that only one type of

19

ligator-detector--matching or complementary--is used in a given FAAST reaction.

Some examples of sequence relationships between adaptor-indexers and ligator detectors are illustrated in Figure 1. Ligator-detectors 1-12 in Figure 1 are designed to match all or part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 13-19 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 21-32 are designed to match all or part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 33-40 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Note that the part of the adjacent sequence of the adaptor-indexer embodied in the ligator-detector is contiguous with the part of the sticky end sequence embodied in the ligator-detector. This is what is meant by adjacent.

Ligator-detectors 1-4 in Figure 1 are designed to match all of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 5-12 are designed to match part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 2-4 and 8-12 are designed to match all or part of the sticky end sequence of adaptor-indexer I and part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 13-16 are designed to be complementary to all of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 17-19 are designed to be complementary to part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 14-16 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer I and part of the adjacent sequence of adaptor-indexer I. Ligator-detector 20 is designed to match all of the recognition sequence of the restriction enzyme (which generates a sticky end compatible with the sticky end of adaptor-indexer I), and all of the adjacent sequence of

20

adaptor-indexer I. Note the extra nucleotide extending beyond the adaptor-indexer sticky end sequence. This is a flanking nucleotide in the recognition sequence.

Ligator-detectors 21-24 are designed to match all of the sticky end
5    sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 25-32 are designed to match part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 22-24 and 28-31 are designed to match all or part of the sticky end sequence of adaptor-indexer II
10   and part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 33-36 are designed to be complementary to all of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 37-40 are designed to be complementary to part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of
15   adaptor-indexer II. Ligator-detectors 34-36 and 40 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer II and part of the adjacent sequence of adaptor-indexer II.

Where the nucleic acid cleaving reagents used in the first digest cleave within the recognition sequence for the nucleic acid cleaving reagent such that
20   the recognition sequence extends beyond the sticky end sequence, the ligator-detector can also match or be complementary to all or part of the recognition sequence. Where the recognition sequence extends beyond the sticky end sequence (for example, six-base recognition sequence and four-base sticky end), the ligator-detector sequence can extend beyond the sticky end sequence of its
25   cognate adaptor-indexer. An example of such a ligator-detector is illustrated in Figure 1 (ligator-detector number 20).

While the ligator-detector can be detected using sequence-based detection systems, the ligator-detectors can also contain a label to facilitate detection of the ligator-detector. Numerous labels are known and can be used
30   for this purpose. Ligator-detectors can also contain or be associated with capture tags to facilitate immobilization or capture of the ligator-detectors. Ligator-detectors can also contain or be associated with sorting tags to facilitate

21

sorting or separation of the ligator-detectors. Ligator-detectors can also be immobilized on a substrate.

Ligator-detectors can also include one or more photocleavable nucleotides to facilitate release of ligator-detector sequences for detection.

5      Photocleavable nucleotides and their use are described in WO 00/04036.

Ligator-detectors need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the ligator-detector have the general structure described herein and be capable of the

10     interactions and reactions required in the disclosed method.

**Detector Probes**

Detector probes are molecules, preferably oligonucleotides, that can hybridize to nucleic acids in a sequence-specific manner. In the disclosed method, detector probes are used to capture ligator-detectors based on

15     complementary sequences present in sample nucleic acid fragments to which the ligator-detectors are hybridized. Detector probes are preferably used in sets having a variety of probe sequences, preferably a set of probes having every possible combination (or hybridizable to every combination) of nucleotide sequence the length of the probe. Detector probes are preferably used in sets

20     where each probe has the same length. Preferred lengths for the probe portion of detector probes are five, six, seven, and eight nucleotides. Detector probes preferably include a probe portion (for hybridization to sample fragments) and linker portions through which the probe portion is coupled to a substrate, capture tag, sorting tag, or label. These linker portions can have any suitable

25     structure and will generally be chosen based on the method of immobilization or synthesis of the detector probes. The linker portion can be made up of or include nucleotides. The linker portions can have any suitable length and preferably are of sufficient length to allow the probe portion to hybridize effectively. For convenience and unless otherwise indicated, reference to the

30     length of detector probes refers to the length of the probe portion of the probes. Immobilized detector probes are detector probes immobilized on a support.

Detector probes can be, and preferably are, immobilized on a substrate. Detector probes can also contain or be associated with capture tags to facilitate immobilization or capture of the probes and ligator-detectors to which they have

been coupled. Detector probes can also contain or be associated with sorting tags to facilitate sorting or separation of the probes and ligator-detectors to which they have been coupled. Detector probes can also contain or be associated with labels to facilitate detection of the probes and ligator-detectors

5      to which they have been coupled.

Detector probes can also include one or more photocleavable nucleotides to facilitate release of probe sequences and ligator-detectors coupled to the probe. Photocleavable nucleotides and their use are described in WO 00/04036.

10      Detector probes need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the probe have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

15  **Probe Arrays**

Different detector probes can be used together as a set. The set can be used as a mixture of all or subsets of the probes, probes used separately in separate reactions, or immobilized in an array. Probes used separately or as mixtures can be physically separable through, for example, the use of capture

20      tags, sorting tags, or immobilization on beads. A probe array (also referred to herein as an array) includes a plurality of probes immobilized at identified or predetermined locations on the array. In this context, plurality of probes refers to a multiple probes each having a different sequence. Each predetermined location on the array has one type of probe (that is, all the probes at that location

25      have the same sequence). Each location will have multiple copies of the probe. The spatial separation of probes of different sequence in the array allows separate detection and identification of ligator-detectors that become coupled to the probes via hybridization of the probes to nucleic acid fragments in a nucleic acid sample. If a ligator-detector is detected at a given location in a probe array,

30      it indicates that the sequence adjacent to the site in the nucleic acid fragment where the ligator-detector hybridized is complementary to the probe immobilized at that location in the array.

24

Adaptor-indexers and ligator-detectors can also be immobilized in arrays. Different modes of the disclosed method can be performed with different components immobilized, labeled, or tagged. Arrays of adaptor-indexers and ligator-detectors can be made and used as described below and

5      elsewhere herein for the detector probes.

Solid-state substrates for use in probe array can include any solid material to which oligonucleotides can be coupled, directly or indirectly. This includes materials such as acrylamide, cellulose, nitrocellulose, glass, polystyrene, polyethylene vinyl acetate, polypropylene, polymethacrylate, polyethylene,

10     polyethylene oxide, glass, polysilicates, polycarbonates, teflon, fluorocarbons, nylon, silicon rubber, polyanhydrides, polyglycolic acid, polylactic acid, polyorthoesters, polypropylfumerate, collagen, glycosaminoglycans, and polyamino acids. Solid-state substrates can have any useful form including thin films or membranes, beads, bottles, dishes, fibers, woven fibers, shaped

15     polymers, particles and microparticles. A preferred form for a solid-state substrate is a microtiter dish. The most preferred form of microtiter dish is the standard 96-well type.

Methods for immobilization of oligonucleotides to solid-state substrates are well established. Detector probes can be coupled to substrates using

20     established coupling methods. For example, suitable attachment methods are described by Pease *et al.*, *Proc. Natl. Acad. Sci. USA* 91(11):5022-5026 (1994), and Khrapko *et al.*, *Mol Biol (Mosk) (USSR)* 25:718-730 (1991). A method for immobilization of 3'-amine oligonucleotides on casein-coated slides is described by Stimpson *et al.*, *Proc. Natl. Acad. Sci. USA* 92:6379-6383 (1995).

25     A preferred method of attaching oligonucleotides to solid-state substrates is described by Guo *et al.*, *Nucleic Acids Res.* 22:5456-5465 (1994).

Methods for producing arrays of oligonucleotides on solid-state substrates are also known. Examples of such techniques are described in U.S. Patent No. 5,871,928 to Fodor et al., U.S. Patent No. 5,54,413, U.S. Patent No.

30     5,429,807, and U.S. Patent No. 5,599,695 to Pease et al.

Although preferred, it is not required that a given probe array be a single unit or structure. The set of probes may be distributed over any number of solid

25

supports. For example, at one extreme, each probe may be immobilized in a separate reaction tube or container.

The probes in arrays can also be designed to have similar hybrid stability. This would make hybridization of fragments to detector probes more 5 efficient and reduce the incidence of mismatch hybridization. The hybrid stability of probes can be calculated using known formulas and principles of thermodynamics (see, for example, Santa Lucia et al., *Biochemistry* 35:3555-3562 (1996); Freier et al., *Proc. Natl. Acad. Sci. USA* 83:9373-9377 (1986); Breslauer et al., *Proc. Natl. Acad. Sci. USA* 83:3746-3750 (1986)). The hybrid 10 stability of the probes can be made more similar (a process that can be referred to as smoothing the hybrid stabilities) by, for example, chemically modifying the probes (Nguyen et al., *Nucleic Acids Res.* 25(15):3059-3065 (1997); Hohsisel, *Nucleic Acids Res.* 24(3):430-432 (1996)). Hybrid stability can also be smoothed by carrying out the hybridization under specialized conditions 15 (Nguyen et al., *Nucleic Acids Res.* 27(6):1492-1498 (1999); Wood et al., *Proc. Natl. Acad. Sci.* USA 82(6):1585-1588 (1985)).

Another means of smoothing hybrid stability of the probes is to vary the length of the probes. This would allow adjustment of the hybrid stability of each probe so that all of the probes had similar hybrid stabilities (to the extent 20 possible). Since the addition or deletion of a single nucleotide from a probe will change the hybrid stability of the probe by a fixed increment, it is understood that the hybrid stabilities of the probes in a probe array will not be equal. For this reason, similarity of hybrid stability as used herein refers to any increase in the similarity of the hybrid stabilities of the probes (or, put another way, any 25 reduction in the differences in hybrid stabilities of the probes). This is useful since any such increased similarity in hybrid stability can improve the efficiency and fidelity of hybridization and ligation of the detector probes.

The efficiency of hybridization and ligation of detector probes to sample fragments can also be improved by grouping detector probes of similar hybrid 30 stability in sections or segments of a probe array that can be subjected to different hybridization conditions. In this way, the hybridization conditions can be optimized for particular classes of probes.

26

**Labels**

To aid in detection and quantitation of ligator-detectors coupled to detector probes, labels can be incorporated into, coupled to, or associated with, ligator-detectors, detector probes, and/or adaptor-indexers. It is preferred that

5   the ligator-detector be labeled. A label is any molecule that can be associated with ligator-detectors, directly or indirectly, and which results in a measurable, detectable signal, either directly or indirectly. A label is associated with a component when it is coupled or bound, either covalently or non-covalently, to the component. A label is coupled to a component when it is covalently

10   coupled to the component. Many suitable labels for incorporation into, coupling to, or association with nucleic acid are known. Examples of labels suitable for use in the disclosed method are radioactive isotopes, fluorescent molecules, phosphorescent molecules, bioluminescent molecules, enzymes, antibodies, and ligands.

15   Examples of suitable fluorescent labels include fluorescein (FITC), 5,6-carboxymethyl fluorescein, Texas red, nitrobenz-2-oxa-1,3-diazol-4-yl (NBD), coumarin, dansyl chloride, rhodamine, 4'-6-diamidino-2-phenylinodole (DAPI), and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. Preferred fluorescent labels are fluorescein (5-carboxyfluorescein-N-hydroxysuccinimide ester) and

20   rhodamine (5,6-tetramethyl rhodamine). Preferred fluorescent labels for simultaneous detection are FITC and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. The absorption and emission maxima, respectively, for these fluors are: FITC (490 nm; 520 nm), Cy3 (554 nm; 568 nm), Cy3.5 (581 nm; 588 nm), Cy5 (652 nm: 672 nm), Cy5.5 (682 nm; 703 nm) and Cy7 (755 nm; 778 nm),

25   thus allowing their simultaneous detection. The fluorescent labels can be obtained from a variety of commercial sources, including Molecular Probes, Eugene, OR and Research Organics, Cleveland, Ohio.

Labeled nucleotides are preferred form of label since they can be directly incorporated into ligator-detectors during synthesis. Examples of labels that can

30   be incorporated into DNA or RNA include nucleotide analogs such as BrdUrd (Hoy and Schimke, *Mutation Research* 290:217-230 (1993)), BrUTP (Wansick *et al.*, J. *Cell Biology* 122:283-293 (1993)) and nucleotides modified with biotin

27

(Langer *et al.*, *Proc. Natl. Acad. Sci. USA* 78:6633 (1981)) or with suitable

haptens such as digoxygenin (Kerkhof, *Anal. Biochem*. 205:359-364 (1992)).

Suitable fluorescence-labeled nucleotides are Fluorescein-isothiocyanate-dUTP,

Cyanine-3-dUTP and Cyanine-5-dUTP (Yu *et al.*, *Nucleic Acids Res.*, 22:3226-

5        3232 (1994)). A preferred nucleotide analog detection label for DNA is BrdUrd

(BUDR triphosphate, Sigma), and a preferred nucleotide analog detection label

for RNA is Biotin-16-uridine-5'-triphosphate (Biotin-16-dUTP, Boehringher

Mannheim). Fluorescein, Cy3, and Cy5 can be linked to dUTP for direct

labeling. Cy3.5 and Cy7 are available as avidin or anti-digoxygenin conjugates

10       for secondary detection of biotin- or digoxygenin-labeled probes.

Labels that are incorporated into nucleic acid, such as biotin, can be

subsequently detected using sensitive methods well-known in the art. For

example, biotin can be detected using streptavidin-alkaline phosphatase

conjugate (Tropix, Inc.), which is bound to the biotin and subsequently detected

15       by chemiluminescence of suitable substrates (for example, chemiluminescent

substrate CSPD: disodium, 3-(4-methoxyspiro-[1,2,-dioxetane-3-2'-(5'-

chloro)tricyclo [3.3.1.1$^{3,7}$]decane]-4-yl) phenyl phosphate; Tropix, Inc.).

Other labels include molecular or metal barcodes, mass labels, and labels

detectable by nuclear magnetic resonance, electron paramagnetic resonance,

20       surface enhanced raman scattering, surface plasmon resonance, fluorescence,

phosphorescence, chemiluminescence, resonance raman, microwave, or a

combination. Mass labels are compounds or moieties that have, or which give

the labeled component, a distinctive mass signature in mass spectroscopy. Mass

labels are useful when mass spectroscopy is used for detection. Preferred mass

25       labels are peptide nucleic acids and carbohydrates. Combinations of labels can

also be useful. For example, color-encoded microbeads having, for example,

265 unique combinations of labels, are useful for distinguishing numerous

components. For example, 256 different ligator-detectors can be uniquely

labeled and detected allowing mutiplexing and automation of the disclosed

30       method.

Useful labels are described in de Haas, R. R., et al., "Platinum porphyrins

as phosphorescent label for time-resolved microscopy," *J. Histochem.*

28

*Cytochem.* 45(9):1279-92 (1997); Karger and Gesteland, "Digital

chemiluminescence imaging of DNA sequencing blots using a charge-coupled

device camera," *Nucleic Acids Res.* 20(24):6657-65 (1992); Keyes, R. S., et al.,

"Overall and internal dynamics of DNA as monitored by five-atom-tethered

5    spin labels," *Biophys. J.* 72(1):282-90 (1997); Kirschstein, S., et al., "Detection

of the DeltaF508 mutation in the CFTR gene by means of time- resolved

fluorescence methods," *Bioelectrochem. Bioenerg.* 48(2):415-21 (1999);

Kricka, L. J., "Selected strategies for improving sensitivity and reliability of

immunoassays," *Clin. Chem.* 40(3):347-57 (1994); Kricka, L. J.,

10   "Chemiluminescent and bioluminescent techniques," *Clin. Chem.* 37(9):1472-

81 (1991); Kumke, M. U., et al., "Temperature and quenching studies of

fluorescence polarization detection of DNA hybridization," *Anal. Chem.*

69(3):500-6 (1997); McCreery, T., "Digoxigenin labeling," *Mol. Biotechnol.*

7(2):121-4 (1997); Mansfield, E. S., et al., "Nucleic acid detection using non-

15   radioactive labeling methods," *Mol. Cell Probes* 9(3):145-56 (1995); Nurmi, J.,

et al., "A new label technology for the detection of specific polymerase chain

reaction products in a closed tube," *Nucleic Acids Res.* 28(8):28 (2000); Oetting,

W. S., et al. "Multiplexed short tandem repeat polymorphisms of the Weber 8A

set of markers using tailed primers and infrared fluorescence detection,"

20   *Electrophoresis* 19(18):3079-83(1998); Roda, A., et al., "Chemiluminescent

imaging of enzyme-labeled probes using an optical microscope-videocamera

luminograph," *Anal. Biochem.* 257(1):53-62 (1998); Siddiqi, A., et al.,

"Evaluation of electrochemiluminescence- and bioluminescence-based assays

for quantitating specific DNA," *J. Clin. Lab. Anal.* 10(6):423-31 (1996);

25   Stevenson, C. L., et al., "Synchronous luminescence: a new detection technique

for multiple fluorescent probes used for DNA sequencing," *Biotechniques*

16(6):1104-11 (1994); Vo-Dinh, T., et al., "Surface-enhanced Raman gene

probes," *Anal. Chem.* 66(20):3379-83 (1994); Volkers, H. H., et al.,

"Microwave label detection technique for DNA in situ hybridization," *Eur. J.*

30   *Morphol.* 29(1):59-62 (1991).

       Metal barcodes, a form of molecular barcode, are 30-300 nm diameter by

400-4000 nm multilayer multi metal rods. These rods are constructed by

electrodeposition into an alumina mold, then the alumina is removed leaving these small multilayer objects behind. The system can have up to 12 zones encoded, in up to 7 different metals, where the metals have different reflectivity and thus appear lighter or darker in an optical microscope depending on the

5      metal; this leads to practically unlimited identification codes. The metal bars can be coated with glass or other material, and probes attached to the glass using methods commonly known in the art; assay readout is by fluorescence from the target, and the identity of the probe is from the light dark pattern of the barcode.

10      Methods for detecting and measuring signals generated by labels are known. For example, radioactive isotopes can be detected by scintillation counting or direct visualization; fluorescent molecules can be detected with fluorescent spectrophotometers; phosphorescent molecules can be detected with a spectrophotometer or directly visualized with a camera; enzymes can be

15      detected by detection or visualization of the product of a reaction catalyzed by the enzyme; antibodies can be detected by detecting a secondary detection label coupled to the antibody. Such methods can be used directly in the disclosed method of amplification and detection. As used herein, detection molecules are molecules which interact with amplified nucleic acid and to which one or more

20      detection labels are coupled. In another form of detection, labels can be distinguished temporally via different fluorescent, phosphorescent, or chemiluminescent emission lifetimes. Multiplexed time-dependent detection is described in Squire et al., J. Microscopy 197(2):136-149 (2000), and WO 00/08443.

25      Quantitative measurement of the amount or intensity of a label can be used. For example, quantitation can be used to determine if a given label, and thus the labeled component, is present at a threshold level or amount. A threshold level or amount is any desired level or amount of signal and can be chosen to suit the needs of the particular form of the method being performed.

30      **Second Adaptors**

Second adaptors are double-stranded nucleic acids containing a single-stranded portion and a double-stranded portion. The single-stranded portion is

30

at one end of the second adaptor and constitutes a sticky end. It is preferable

that the protruding single strand (sticky end) have two, three, four, or five

nucleotides. The double-stranded portion of second adaptor may have any

convenient sequence or length. In general, the sequence and length of the

5    double-stranded portion is selected to be adapted to subsequent steps in the

method. For example, sequences in the second adaptor may be used for primer

or probe hybridization. If the fragments in the samples to which adaptors have

been ligated are to be amplified, the second adaptors can provide sequence for

primer hybridization. Thus, preferred sequence composition and length for the

10   double-stranded portion of second adaptors will generally be those that are

useful for primer hybridization.

Generally, the sequence of the double-stranded portion of a second

adaptor should not include the recognition sequence of any nucleic acid

cleaving reagent to be used in a subsequent step in the method. It is preferred

15   that second adaptors not have any sequences that are self-complementary. It is

considered that this condition is met if there are no complementary regions

greater than six nucleotides long without a mismatch or gap.

A set of second adaptors for use in the disclosed method can include

different second adaptors where the single-stranded portion each have a

20   different nucleotide sequence compatible with a sticky end sequence generated

by one of the second restriction enzymes. It is preferable that the members of a

set of second adaptors contain a double-stranded portion which is identical for

each member of the set.

Second adaptors can also contain or be associated with capture tags to

25   facilitate immobilization or capture of fragments to which second adaptors have

been coupled. Second adaptors can also contain or be associated with sorting

tags to facilitate sorting or separation of fragments to which second adaptors

have been coupled. Second adaptors can also contain or be associated with

labels to facilitate detection of fragments to which second adaptors have been

30   coupled. Second adaptors can also be immobilized on a substrate.

31

## Capture Tags

A capture tag is any compound that can be used to separate compounds or complexes having the capture tag from those that do not. Preferably, a capture tag is a compound, such as a ligand or hapten, that binds to or interacts

5 with another compound, such as ligand-binding molecule or an antibody. It is also preferred that such interaction between the capture tag and the capturing component be a specific interaction, such as between a hapten and an antibody or a ligand and a ligand-binding molecule.

Preferred capture tags, described in the context of nucleic acid probes,

10 are described by Syvnen *et al.*, *Nucleic Acids Res.*, 14:5037 (1986). Preferred capture tags include biotin, which can be incorporated into nucleic acids. In the disclosed method, capture tags incorporated into adaptor-indexers or second adaptors can allow sample fragments (to which the adaptors have been coupled) to be captured by, adhered to, or coupled to a substrate. Such capture allows

15 simplified washing and handling of the fragments, and allows automation of all or part of the method.

Capturing sample fragments on a substrate may be accomplished in several ways. In one embodiment, capture docks are adhered or coupled to the substrate. Capture docks are compounds or moieties that mediate adherence of

20 a sample fragment by binding to, or interacting with, a capture tag on the fragment. Capture docks immobilized on a substrate allow capture of the fragment on the substrate. Such capture provides a convenient means of washing away reaction components that might interfere with subsequent steps.

Substrates for use in the disclosed method can include any solid

25 material to which components of the assay can be adhered or coupled. Examples of substrates include, but are not limited to, materials such as acrylamide, cellulose, nitrocellulose, glass, polystyrene, polyethylene vinyl acetate, polypropylene, polymethacrylate, polyethylene, polyethylene oxide, polysilicates, polycarbonates, teflon, fluorocarbons, nylon, silicon rubber,

30 polyanhydrides, polyglycolic acid, polylactic acid, polyorthoesters, polypropylfumerate, collagen, glycosaminoglycans, and polyamino acids.

32

Substrates can have any useful form including thin films or membranes, beads, bottles, dishes, fibers, woven fibers, shaped polymers, particles and microparticles. Preferred forms of substrates are plates and beads. The most preferred form of beads are magnetic beads.

5          In one embodiment, the capture dock is an oligonucleotide. Methods for immobilizing and coupling oligonucleotides to substrates are well established. For example, suitable attachment methods are described by Pease *et al.*, *Proc. Natl. Acad. Sci. USA* 91(11):5022-5026 (1994), and Khrapko *et al.*, *Mol Biol (Mosk) (USSR)* 25:718-730 (1991). A method for immobilization of

10    3'-amine oligonucleotides on casein-coated slides is described by Stimpson *et al.*, *Proc. Natl. Acad. Sci. USA* 92:6379-6383 (1995). A preferred method of attaching oligonucleotides to solid-state substrates is described by Guo *et al.*, *Nucleic Acids Res.* 22:5456-5465 (1994).

          In another embodiment, the capture dock is the anti-hybrid antibody.

15    Methods for immobilizing antibodies to substrates are well established. Immobilization can be accomplished by attachment, for example, to aminated surfaces, carboxylated surfaces or hydroxylated surfaces using standard immobilization chemistries. Examples of attachment agents are cyanogen bromide, succinimide, aldehydes, tosyl chloride, avidin-biotin,

20    photocrosslinkable agents, epoxides and maleimides. A preferred attachment agent is glutaraldehyde. These and other attachment agents, as well as methods for their use in attachment, are described in *Protein immobilization: fundamentals and applications*, Richard F. Taylor, ed. (M. Dekker, New York, 1991), Johnstone and Thorpe, *Immunochemistry In Practice* (Blackwell

25    Scientific Publications, Oxford, England, 1987) pages 209-216 and 241-242, and *Immobilized Affinity Ligands*, Craig T. Hermanson *et al.*, eds. (Academic Press, New York, 1992). Antibodies can be attached to a substrate by chemically cross-linking a free amino group on the antibody to reactive side groups present within the substrate. For example, antibodies may be chemically

30    cross-linked to a substrate that contains free amino or carboxyl groups using glutaraldehyde or carbodiimides as cross-linker agents. In this method, aqueous solutions containing free antibodies are incubated with the solid-state substrate

33

in the presence of glutaraldehyde or carbodiimide. For crosslinking with
glutaraldehyde the reactants can be incubated with 2% glutaraldehyde by
volume in a buffered solution such as 0.1 M sodium cacodylate at pH 7.4.
Other standard immobilization chemistries are known by those of skill in the

5      art.

**Sorting Tags**

A sorting tag is any compound that can be used to sort or separate
compounds or complexes having the sorting tag from those that do not. In
general, all capture tags can be a sorting tag. Sorting tags also include

10     compounds and moieties that can be detected and which can mediate the sorting
of tagged components. Such forms of sorting tags are generally not also capture
tags. For example, a fluorescent moiety can allow sorting of components
tagged with the moiety from those that are not (or those with a different tag).
However, such a fluorescent moiety does not necessarily have a suitable capture

15     dock with which it can interact and be captured. Preferably, a sorting tag is a
label, such as a fluorescent label, that can mediate sorting.

**Amplification Target Circle**

An amplification target circle (ATC) is a circular single-stranded DNA
molecule, generally containing between 40 to 1000 nucleotides, preferably

20     between about 50 to 150 nucleotides, and most preferably between about 50 to
100 nucleotides. Portions of ATCs have specific functions making the ATC
useful for rolling circle amplification (RCA). These portions are referred to as
the primer complement portion and the reporter tag portions. The primer
complement portion and the reporter tag portion are required elements of an

25     amplification target circle. Those segments of the ATC that do not correspond
to a specific portion of the ATC can be arbitrarily chosen sequences. It is
preferred that ATCs do not have any sequences that are self-complementary. It
is considered that this condition is met if there are no complementary regions
greater than six nucleotides long without a mismatch or gap.

30     An amplification target circle, when replicated, gives rise to a long
DNA molecule containing multiple repeats of sequences complementary to the
amplification target circle. This long DNA molecule is referred to herein as

tandem sequences DNA (TS-DNA). TS-DNA contains sequences complementary to the primer complement portion and the reporter tag portions. These sequences in the TS-DNA are referred to as primer sequences (which match the sequence of the rolling circle replication primer) and reporter tags.

5      Amplification target circles and their use are further described in U.S. Patent No. 5,854,033.

## Method

The FAAST method involves the following basic steps. A nucleic acid sample, embodied in double stranded DNA, is incubated with one or more

10    nucleic acid cleaving reagent, preferably restriction endonucleases, such that a set of DNA fragments having sticky ends with a variety of sequences is generated. Preferred for this purpose is the use of a single type II-S restriction endonuclease having an offset cleavage site. Since such class II-S restriction endonucleases cleave at a site different from the recognition sequence, this

15    results in a set of DNA fragments having sticky ends with a variety of sequences. A similar effect can be obtained by digesting the nucleic acid sample with a mixture of restriction endonucleases or other nucleic acid cleaving reagents which cleave at their recognition site.

For a four base sticky end, there are 256 possible sequences. The

20    general formula is $N = 4^X$ where X is the length of the sticky end and N is the number of possible sequences. In a sufficiently complex nucleic acid sample, all of these sequences will be represented in the ends of the set of DNA fragments. The nucleic acid sample is also divided into aliquots (referred to as index samples); preferably as many aliquots as there are sticky end sequences

25    (that is, $N = 4^X$ aliquots). Where multiple restriction endonucleases are used, the nucleic acid sample is preferably divided into index samples before digestion. Where a single restriction endonuclease is used, the nucleic acid sample is preferably divided into index samples following digestion. Each index sample is then mixed with a different adaptor-indexer, each of which has

30    a sticky end compatible with one of the possible sticky ends on the DNA fragments in that index sample. The adaptor-indexes are then coupled onto

35

compatible DNA fragments. The basic method then proceeds to hybridization of the nucleic acid sample with ligator-detectors.

Optionally, prior to hybridization with ligator-detectors, each index sample can then be incubated with one or more other nucleic acid cleaving

5    reagents (referred to as second nucleic acid cleaving reagents), preferably restriction enzymes. Second nucleic acid cleaving reagents are preferably restriction enzymes having a four base recognition sequences. All index samples are preferably digested with the same nucleic acid cleaving reagent(s). Alternatively, the index samples can be further divided into secondary index

10   samples each of which is digested with a different second nucleic acid cleaving reagent or set of nucleic acid cleaving reagents. A second adaptor can then be coupled to the DNA fragments in the index samples (or secondary index samples). Preferably, the same second adaptor is used for each index sample. Different second adaptors are preferably used with secondary index sample

15   derived from the same index sample. In this case, it is preferred that the same set of second adaptors be used with each set of secondary index samples. DNA fragments in each index sample (or secondary index sample) now have adaptors ligated to each end. The DNA fragments can then be amplified using any suitable method, such as PCR. Sequences in the adaptors can be used as primer

20   binding sites for this amplification.

Optionally, prior to hybridization with ligator-detectors, the index samples (or secondary index samples) can divided into further aliquots. These are referred to as restricted index samples and non-restricted index samples (or restricted secondary index samples and non-restricted secondary index samples,

25   if there are secondary index samples). Generally, the index samples (or secondary index samples) can be divided into one or more restricted index samples and one non-restricted index sample. The restricted index samples (or restricted secondary index samples), but not the non-restricted index sample (or non-restricted secondary index sample) are then each incubated with a different

30   nucleic acid cleaving reagents (referred to as third nucleic acid cleaving reagents), preferably restriction enzymes. The third nucleic acid cleaving reagents are preferably different from any of the nucleic acid cleaving reagents

36

or second nucleic acid cleaving reagents with which the sample has been digested.

In some cases, the third nucleic acid cleaving reagents will cleave some DNA fragments in the restricted index samples (or restricted secondary index samples), thus making the fragment incompetent for amplification (another optional step prior to hybridization of the nucleic acid sample with ligator-detectors). In this way, the signals generated by the restricted index samples and non-restricted index sample (or restricted and non-restricted secondary index samples) can differ, and fragments containing the recognition sequence of one of the third restriction enzymes can be identified.

Secondary index samples, restricted index samples, non-restricted index samples, restricted secondary index samples, and non-restricted secondary index samples are referred to collectively herein as derivative index samples. Each is derived from an index sample and, in some cases, from another derivative index sample. In general, only those derivative index samples last generated are carried forward in the method. For example, if secondary index samples are created, the original index samples from which they were derived are no longer carried forward in the method (the secondary index samples are). Similarly, if restricted and non-restricted secondary index samples are created, then neither the original index samples nor the secondary index samples from which the restricted and non-restricted secondary index samples were derived are carried forward in the method. However, additional information may be gained by carrying forward all or some of the index samples and derivative index samples.

After the basic steps (and any desired optional steps), the DNA fragments in each index sample (or each derivative index sample, if used) are hybridized to ligator-detectors. One portion of each ligator-detector matches or is complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers. Preferably, the ligator-detector has sequence matching or complementary to one of the possible sticky end sequences generated by digestion with the restriction enzymes. The ligator-detector can also be complementary to nucleotides in the recognition sequence of the restriction enzymes if restriction enzymes that cleave in their recognition

sequence are used. Ligator-detectors can also have sequence matching or
complementary to common sequences in the adaptor-indexers. In this case, the
appropriate ligator-detector is used with each of the index samples (or each
derivative index sample, if used). That is, the ligator-detector oligonucleotide

5      used in each index sample (or each derivative index sample) matches or is
complementary to sequence, including sticky end sequence (and recognition
sequence, if overlapping with the sticky end sequence), in the adaptor-indexer
sequence used in that index sample (or used in the index sample from which the
derivative index sample was derived). Alternatively, the ligator-detector can

10     have sequence matching or complementary to sequence of the nucleic acid
fragment (to which an adaptor-indexer has been coupled) adjacent to the sticky
end sequence and on opposite the side of the fragment from the adaptor-indexer.
Such ligator-detectors are preferred for use in VAAST.

       Finally, each index sample (or each derivative index sample) is mixed

15     with detector probes and the probes are coupled to the ligator-detectors.
Preferably, the set of probes used include every possible sequence of a given
length (for example, every possible six base sequence). The detector probes can
be immobilized in an array.

       The ends of the detector probes and the ligator-detectors are coupled

20     together only if the probe hybridizes adjacent to the end of the ligator-adaptor.
Thus, a ligator-detector is coupled to a detector probe only when a sequence
complementary to the probe is present immediately adjacent to the region in a
DNA fragment from the original sample to which the end of the ligator-detector
hybridizes (preferably the sticky end sequences). Examples of the relationship

25     and interaction of various components of the disclosed method are illustrated in
Figures 2, 3, 4, and 5.

       Each processed DNA fragment (that is, each DNA fragment to which an
adaptor-indexer was ligated) from the sample will result in an association of a
ligator-detector (and an adaptor-indexer) with a detector probe. This

30     association will be detected through a signal generated from one of the
associated components. In a preferred form of the disclosed method, the set of
detector probes (if multiple sets are used) in which the signal for a given

38

fragment is determined by the sequence of the original sticky end sequence (or recognition sequence). Each different sticky end or recognition sequence is processed in a separate index sample; a separate set of detector probes is used for each index sample or derivative index sample. The probe in the set of

5      probes to which the signal for a given fragment is associated and detected is determined by the sequence in the DNA fragment adjacent to the sticky end sequence (or recognition sequence) since the detector probe must hybridize to this sequence in order to be coupled to the ligator-detector hybridized to the DNA fragment. A complex nucleic acid sample will produce a unique pattern

10     of signals in the probe sets. It is this pattern that allows unique cataloging of nucleic acid samples and sensitive and powerful comparisons of the patterns of signals produced from different nucleic acid samples.

The set of detector probes, and particular probe in the set, in which a signal is associated identifies the sequence of the sticky end of the DNA

15     fragment that gave rise to the signal and of the sequence adjacent to the sticky end (or the recognition sequence of the restriction enzyme and of the sequence adjacent to the recognition sequence). This is a ten base sequence when a four base sticky end and six base probes are used. The set of detector probes identifies the sticky end sequence (the four bases) because each set of detector

20     probes used a different adaptor-indexer having a different sticky end sequence. The particular probe identifies the sequence adjacent to the sticky end (the six bases) because each probe has a different probe with a different sequence. Only the probe with sequence complementary to the adjacent sequence will hybridize and thus become associated with a signal. The fixed relationship between the

25     recognition sequence and the cleavage site of a type II-S restriction enzyme, when used, and the identity of the recognition sequence, provide additional sequence information about the DNA fragment.

The information generated is similar when probe arrays are used. The array (if multiple arrays are used) in which the signal for a given fragment is

30     detected is determined by the sequence of the original sticky end sequence (or recognition sequence). Each different sticky end or recognition sequence is processed in a separate index sample; a separate array is used for each index

39

sample or derivative index sample. The location in the array in which the signal
for a given fragment is detected is determined by the sequence in the DNA
fragment adjacent to the sticky end sequence (or recognition sequence) since the
probe must hybridize to this sequence in order to be coupled to the ligator-
5       detector hybridized to the DNA fragment. A complex nucleic acid sample will
produce a unique pattern of signals on the arrays.

The array, and location in the array, where a DNA fragment generates a
signal identifies the sequence of the sticky end of the DNA fragment and of the
sequence adjacent to the sticky end (or the recognition sequence of the
10      restriction enzyme and of the sequence adjacent to the recognition sequence).
This is a ten base sequence when a four base sticky end and six base probes are
used. The array identifies the sticky end sequence (the four bases) because each
array used a different adaptor-indexer having a different sticky end sequence.
The location in the array identifies the sequence adjacent to the sticky end (the
15      six bases) because each location in the array has a different probe with a
different sequence. Only the probe with sequence complementary to the
adjacent sequence will hybridize and thus become associated with a signal. The
fixed relationship between the recognition sequence and the cleavage site of a
type II-S restriction enzyme, when used, and the identity of the recognition
20      sequence, provide additional sequence information about the DNA fragment.

The disclosed method is performed using one or more nucleic acid
cleaving reagents that collectively produce a plurality of different sticky end
sequences. Preferably, the sticky end sequences generated by the nucleic acid
cleaving reagent are not limited by the recognition sequence of the nucleic acid
25      cleaving reagent. The sticky ends generated are preferably 2, 3, 4 or 5
nucleotides long. Preferred nucleic acid cleaving reagents for use in the
disclosed method are type II-S restriction endonucleases, which are enzymes
that cleave DNA at locations outside of (or offset from) the recognition site and
which generate sticky ends. Examples of Type II-S restriction endonucleases
30      are FokI, BbvI, HgaI, BspMI and SfaNI.

Nucleic acid cleaving reagents for use in the disclosed method produce
sticky ends encompassing permutations and combinations of the four

40

nucleotides, A, C, G, and T. The larger the number of protruding bases, the greater the number of possible permutations and combinations of terminal nucleotide sequences, and the more specific the indexing is likely to be. For example, a restriction endonuclease such as FokI, which releases fragments with

5    four base, 5'-protruding sticky ends, will generate fragments having $4^4$ or 256 possible protruding tetranucleotide ends. Cleavage of a cDNA sample having an average of 12,000 different cDNAs with the restriction endonuclease FokI will produce a mixture of fragments with four base, 5'-protruding ends. For a sample containing a random distribution of bases, on average, FokI cuts twice in

10    every $4^5$ base pairs giving an average fragment size of 512 base pairs. If the average length of cDNA is 1,700 base pairs, each cDNA will produce approximately four fragments. The entire sample will contain approximately 4 * 12,000 = 48,000 fragments. There are $4^4 = 256$ possible tetranucleotide sequences and therefore 256 possible identities for each sticky end. On average,

15    there will be 48,000/256 = 188 fragments with a given sticky end sequence. Each of these fragments is sorted by hybridization to different probes based on the sequence adjacent to the sticky end sequence in each fragment. A hexamer probe array has 4,096 different six nucleotide probes. Thus, only 188 of the 4,096 hexamers in the probe array will ligate to a ligator-detector, on average.

20    With 256 probe arrays each having 4,096 different hexamer detector probes, there are 256 * 4,096 = 1,048,576 "bins" in which to distribute 48,000 fragments. This leaves ample opportunity to identify different patterns when using different cDNA samples.

Cleavage of human genomic DNA (which has a haploid number of 3 X

25    $10^9$ base pairs) with the restriction endonuclease Bsp24I will release a large and complex mixture of fragments with five base, 3'-protruding ends. On average, Bsp24I cuts twice in every $4^6$ base pairs giving an average fragment size of 2048 base pairs, and resulting in 3 X $10^9$/2048 = approximately 1.5 X $10^6$ fragments. There are $4^5 = 1024$ possible pentanucleotide sequences and

30    therefore 1024 possible identities for each sticky end. On average, there will be 1.5 X $10^6$/1024 = 1,465 fragments with a given sticky end sequence. Each of these fragments is sorted by hybridization to different probes based on the

41

sequence adjacent to the sticky end sequence in each fragment. An heptamer probe array has 16,384 different seven nucleotide probes. Thus, only 1,465 of the 16,384 heptamers in the array will ligate to a ligator-detector, on average. With 1024 detector arrays each having 16,384 different heptamer probes, there

5 are $1024 * 16,384 = 1.6 \times 10^7$ "bins" in which to distribute $1.5 \times 10^6$ fragments.

Cleavage of a cDNA sample with twenty different restriction endonucleases having six-base recognition sequences will produce a mixture of fragments with sticky ends. On average, restriction endonucleases having six-

10 base recognition sequences cut once every $4^6 = 4096$ base pairs. If the sample contains approximately 12,000 cDNA molecules with an average length of cDNA of 1,500 base pairs, cleavage with one of the restriction enzymes will result about 3200 cuts (and thus 6400 DNA fragments with sticky ends). Further cleavage of the sample (second digest) with two different restriction

15 endonucleases having four-base recognition sequences will result in additional cuts once every $4^4 = 256$ base pairs. Since the second digest will, in many cases, result in cuts on each fragment, this will result in (for each of the 20 * 2 = 40 secondary index samples) approximately 6,400 fragments, each approximately 256 base pairs long.

20 If five different restriction endonucleases having four-base recognition sequences are used for the third digest, approximately half of the fragments in each restricted secondary index sample will be cleaved (since these restriction enzymes will cut about once every 256 base pairs). Thus, there will be approximately 3,200 fragments (intact, with both an adaptor-indexer and a

25 second adaptor) in each of the 20 * 2 * 5 = 200 restricted secondary index samples (there will be approximately 6,400 fragments in the non-restricted secondary index sample). Each of these fragments is sorted by hybridization to different probes based on the sequence adjacent to the sticky end sequence in each fragment. A hexamer probe array has 4,096 different six nucleotide

30 probes. Thus, only 3,200 of the 4,096 hexamers in the array will ligate to a ligator-detector, on average. With 200 probe arrays each having 4,096 different hexamer probes, there are 200 * 4,096 = 819,200 "bins" in which to distribute

42

the of 3,200 * 200 = 640,000 total fragments (a heptamer array would provide 200 * 16,384 = 3,276,800 "bins").

As these examples illustrate, the length of the recognition sequence, the length of the sticky end generated, and the length of the probes used together determine the number of data bins into which the nucleic acid fragments are sorted. By using sticky ends and detector probes of sufficient length, the sorting of fragments can be matched to the complexity of the sample being analyzed.

The use of a comprehensive panel of adaptor-indexers provides a means for attaching specific functional modifications to selected subsets of a complex mixture of nucleic acid fragments and identifying the molecules so modified. Such a defined subset of molecules may be further resolved by additional cleavage and indexing, or by any of the established techniques such as cloning, PCR amplification, or gel electrophoresis. Individual members of the class may be distinguished by identifying characteristics such as length, sequence, or restriction endonuclease maps. The sequence of the sticky ends of the adaptor-indexers provides a means of indexing a large number of nucleic acid fragments.

Detector probes of different sequence can be immobilized at different locations on a probe array. In this way, the sequence of the probes on the probe array and the sequence of nucleic acid fragments in the index samples determine where on the array ligator-detectors become coupled. The presence, amount, presence and amount, or absence of ligator-detector at different locations in the probe arrays thus forms a pattern of signals that provides a signature or fingerprint of a nucleic acid sample based on the presence or absence of specific nucleic acid sequences in the sample. For this reason, cataloging of this pattern of signals (that is, the pattern of the presence, amount, presence and amount, or absence of ligator-detectors) is an embodiment of the disclosed method that is of particular interest. As discussed elsewhere herein, the probability of probe mismatch can be used to create more complex catalogs based on differential hybridization of particular fragments to different detector probes.

Catalogs can be made up of, or be referred to, as, for example, a pattern of ligator-detectors on probe arrays, a pattern of the presence of ligator-

43

detectors on probe arrays, a catalog of nucleic acid fragments in a sample, or a catalog of nucleic acid sequences in a sample. The information in the catalog is preferably in the form of positional information (that is, location in the detector array) or, more preferably, in the form of sequences. Preferred sequence

5    information for catalogs include sequences of detector probes to which a ligator-detector was coupled and sequences of nucleic acid fragments present in the sample (derived from the locations in the detector array where ligator-detectors were coupled).

Such catalogs of nucleic acid samples can be compared to a similar

10    catalog derived from any other sample to detect similarities and differences in the samples (which is indicative of similarities and differences in the nucleic acids in the samples). For example, a catalog of a first nucleic acid sample can be compared to a catalog of a sample from the same type of organism as the first nucleic acid sample, a sample from the same type of tissue as the first

15    nucleic acid sample, a sample from the same organism as the first nucleic acid sample, a sample obtained from the same source but at a different time than the first nucleic acid sample, a sample from a different organism than the first nucleic acid sample, a sample from a different type of tissue than the first nucleic acid sample, a sample from a different strain of organism than the first

20    nucleic acid sample, a sample from a different species of organism than the first nucleic acid sample, or a sample from a different type of organism than the first nucleic acid sample.

The same type of tissue is tissue of the same type such as liver tissue, muscle tissue, or skin (which may be from the same or a different organism or

25    type of organism). The same organism refers to the same individual, animal, or cell. For example, two samples taken from a patient are from the same organism. The same source is similar but broader, referring to samples from, for example, the same organism, the same tissue from the same organism, or the same cDNA library. Samples from the same source that are to be compared are

30    preferably collected at different times (thus allowing for potential changes over time to be detected). This is especially useful when the effect of a treatment or change in condition is to be assessed. A different organism refers a different

44

individual organism, such as a different patient, a different individual animal. Different organism includes a different organism of the same type or organisms of different types. A different type of organism refers to organisms of different types such as a dog and cat, a human and a mouse, or *E. coli* and *Salmonella*. A

5   different type of tissue refers to tissues of different types such as liver and kidney, or skin and brain. A different strain or species of organism refers to oganisms differing in their species or strain designation as those terms are understood in the art.

The presence, amount, presence and amount, or absence of ligator-

10   detectors coupled to detector probes can be accomplished by detection of labels incorporated into, coupled to, or associated with the ligator-detectors. Alternatively, the ligator-detectors can be detected based on detection of their sequence. These detections are generally referred to as direct detection of coupling of ligator-detectors. Any of the numerous sequence-specific detection

15   techniques can be used for this purpose, including, for example, hybridization of labeled probes. The presence, amount, presence and amount, or absence of ligator detectors can also be detected by generating a signal mediated by the ligator-detector. Use of the ligator-detector as a primer for rolling circle replication, described below, is a preferred example of this. The presence,

20   amount, presence and amount, or absence of ligator detectors can also be detected by detecting the detector probe to which the ligator-detector is coupled, the adaptor-indexer associated with the coupled ligator-detector, or both. These detections are generally referred to as indirect detection of coupling of ligator-detectors.

25   The signal to be detected for the nucleic acid fragments can be increased by nucleic acid amplification during the method. It is preferred either that the nucleic acid fragments to which adaptor-indexers have been coupled (referred to as adaptor/fragments) be amplified or that the ligator-detectors that have been coupled to detector probes be amplified or mediate amplification of another

30   nucleic acid. In the first case, the adaptor/fragments can be amplified using any suitable method. These include polymerase chain reaction (PCR), ligase chain reaction (LCR), self-sustained sequence replication (3SR), nucleic acid

45

sequence based amplification (NASBA), strand displacement amplification (SDA), amplification with Qβ replicase, and rolling circle amplification (RCA) (Birkenmeyer and Mushahwar, *J. Virological Methods*, 35:117-126 (1991); Landegren, *Trends Genetics*, 9:199-202 (1993); Lizardi *et al.*, *Nature Genetics*

5      19(3):225-232 (1998)). In the case of ligator-detector amplification, a preferred form of amplification is rolling circle amplification of a single-stranded circular DNA molecule primed by the ligator-detector. In this way, a long tandem repeat of the DNA circle is generated with the amplified strand anchored to the detector array via the ligator-detector. This technique of amplification is

10     described in PCT application WO 97/19193. If the ligator-detector is used as a rolling circle replication primer, there is no need to incorporate a label in the ligator-detector since the amplified DNA can be detected (either directly or via an incorporated label).

Amplification of the adaptor/fragments is facilitated by the presence of

15     adaptor-indexer sequence at the end of the fragment (and by the presence of second adaptor sequence at the other end, if a second adaptor is used). For example, the adaptor sequences can be used for amplification primer sequences. The adaptor sequences can also be used to circularize the adaptor/fragments for subsequent amplification by rolling circle replication. Rolling circle

20     amplification is described in U.S. Patent No. 5,854,033 and PCT application WO 97/19193.

In another embodiment, the strands of nucleic acid fragments to which adaptor-indexers have been coupled (adaptor/fragments) can be separated prior to hybridization to the ligator-detectors. Such strand separation can improve the

25     efficiency of ligator-detector hybridization. This separation can be accomplished using any suitable technique. Strand separation is preferably accomplished by including a capture tag or sorting tag on one of the strands of the adaptor-indexers. Such a capture tag can then be used to immobilize one strand of the adaptor/fragments while the other strands are washed away. Either

30     the immobilized or washed strand can be carried forward in the method. A sorting tag can allow separation of the strands by a sorting technique.

46

In another embodiment, the concentration of the various nucleic acid fragments in the index samples are normalized. Normalization can be preformed either before after any amplification step that may be used. A preferred technique for fragment normalization involves immobilizing one

5    strand of the nucleic acid fragments, denaturing the nucleic acid fragments, renaturing the nucleic acid fragments for a time greater than the $c_0t_{1/2}$ for abundant nucleic acid fragments and less than the $c_0t_{1/2}$ for rare nucleic acid fragments, and collecting the un-renatured nucleic acid fragments.

To correct hybridization errors one may use one of two general

10   approaches in a singular, dual or sequential manner. For purpose of illustration consider a hybridization and ligation experiment carried out in a glass microarray detection system. In the first approach one treats a hybridized array, prior to the ligation step, with a process that blocks ligation of the indexer-detector to an incorrect probe sequence. The key to this approach is to block

15   (sterically hinder) the nick to be ligated such that the covalent bonding mechanism is unable to make the bond. In practice one would take a hybridized array and contact it with the blocking entity and a suitable buffer and hold the system at a suitable temperature for an appropriate time for the masking entity to become correctly located on the mismatches, then continuing with ligation.

20   A number of single chain antibodies selected to bind to particular mismatches are of use for this method, as are a number of yeast mismatch repair and cell cycle checkpoint proteins such as MSH2-MSH6 (Habraken et al., ATP-dependent assembly of a ternary complex consisting of a DNA mismatch and the yeast MSH2-MSH6 and MLH1-PMS1 protein complexes. J Biol Chem

25   273(16):9837-41 (1998)); MSH2-MSH3 (Habraken et al., Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. Curr Biol 6(9):1185-7 (1996)); MutS; and Mec1 (Paciotti et al., Mec1p is essential for phosphorylation of the yeast DNA damage checkpoint protein Ddc1p, which physically interacts with Mec3p. Embo J

30   17(14):4199-209 (1998)).

The second approach is to treat a covalently bonded probe-target sample with a system that recognizes and specifically cleaves the strand containing the

47

mismatch. This second approach utilizes recognition of mismatches and destructive targeting or 'error checking' of the mismatched region such that in the final stringent washes of the probe array the destroyed mismatched sample is washed away from the chip before the signal detection procedure. In practice

5      one would take the hybridized and ligated probe array and contact it with the error checking reagents in a suitable buffer and hold the system at a suitable temperature for an appropriate time for the error checking reagent to locate the mismatches and destroy the strand containing the mismatch prior to the final stringent washes. One or more of the following 'error checking' systems can be

10     used: T4 endonuclease VII; proteins MutH, MutL, and MutS together; MutY (Lu and Hsu, *Detection of single DNA base mutations with mismatch repair enzymes.* Genomics 14(2):249-55 (1992)); glycosylase (Neddermann and Jiricny, *Efficient removal of uracil from G.U mispairs by the mismatch-specific thymine DNA glycosylase from HeLa cells.* Proc Natl Acad Sci U S A

15     91(5):1642-6 (1994)); mammalian topoisomerase I (Yeh *et al., Mammalian topoisomerase I has base mismatch nicking activity.* J Biol Chem 269(22):15498-504 (1994)); and T7 endonuclease I.

In particular, the fidelity of the hybridization/coupling steps in the disclosed method--the steps where adaptor-indexers are coupled to fragments

20     and where ligator-detectors are coupled to detector probes--can be increased by digesting the samples with a mismatch nuclease such as T4 endonuclease VII. T4 endonuclease VII cleaves double-stranded DNA where there is a mismatch. By "unligating" illegitimate hybridization/ligations, the mis-cataloging of nucleic acid fragments will be made less likely. Fidelity of

25     hybridization/ligation steps can also be maximized by using appropriately stringent hybridization conditions.

The following provides examples of FAAST and the sequence and sequence relationships that can be obtained. When a single type II-S restriction enzyme is used in the first digest, the sequence information obtainable can be

30     illustrated with the following structures:

```
DNA fragment:
..NNNNXXXX..NNNNRRRROOOOOOOOOOSSSSNNNN..
Sequence information:              RRRRROOOIIIIIISSSS
```

In these structures, each character represents a nucleotide. N represents any nucleotide (having no special identity or relationship to the method). R represents a nucleotide in the recognition sequence of the type II-S restriction enzyme. O represents a nucleotide in the offset between the recognition site and

5    the cleavage site of the type II-S restriction enzyme. S represents a nucleotide in the sticky end resulting from cleavage with the type II-S restriction enzyme. X represents a nucleotide in the recognition/cleavage site of the second restriction enzyme. I represents a nucleotide complementary to the immobilized oligonucleotide.

10    From the DNA fragment ...NNNNXXXX...NNNNRRRRROOOOOO OOOSSSSNNNN..., the sequence information RRRRROOOIIIIISSSS can be obtained. In this example, the type II-S restriction enzyme has a five base recognition sequence, a nine base offset to the cleavage site (that is, there are nine bases between the recognition sequence and the cleavage site), and creates

15    a four base sticky end. The probes contain hexamer sequences. Each array location where a signal is generated in this example thus represents a specific sequence : nnnnn---nnnnnnnnnn (where n represents an identified nucleotide and each - represents an unidentified nucleotide). This is referred to as a determined sequence. The portion of the nucleic acid fragments for which the

20    sequence is determined corresponds to the sticky end sequence, the sequence adjacent to the sticky end sequence to which the detector probe hybridized, and the recognition sequence of the restriction enzyme (S, I, and R, respectively).

This sequence information can also be represented by the structure A-B-C-D

25    where A is the recognition sequence of the restriction enzyme, B is the gap of unknown sequence, C is the sequence to which the immobilized oligonucleotide hybridized, and D is the sticky end sequence. The gap represents the nucleotides between the recognition sequence and the sequence to which the probe hybridized. C is always adjacent to the sticky end sequence D.

30    In the example above, A is RRRRR, B is OOO, C is IIIIII, and D is SSSS. Cleavage by the second restriction enzyme also adds useful information since without cleavage there would be no amplification (and thus no further

49

processing of that nucleic acid molecule) with the result that a signal will be absent.

The sequence information that can be obtained with the disclosed method can be further illustrated using a specific example of a nucleic acid

5 fragment. Assume a nucleic acid sample containing a nucleic acid fragment with the sequence

```
..CGGTGGATGACTTGAAGCTATGCTTAGG..
..GCCACCTACTGAACTTCGATACGAATCC..
```

If the sample is digested with FokI--a type II-S restriction enzyme with a

10 recognition sequence of GGATG and a cleavage site offset by 9 and 13 nucleotides--the fragment will be cleaved to generate the following fragments (the FokI recognition sequence is shown in bold)

```
..CGGTGGATGACTTGAAGC          TATGCTTAGG..
..GCCACCTACTGAACTTCGATAC          GAATCC..
```

15 When the corresponding adaptor-indexer is ligated to fragment and the ligated fragment is hybridized to the corresponding ligator-detector, the following nucleic acid is obtained (sequence from the adaptor-indexer is underlined)

```
                         TATGCGGTATTACAGCCTATTGGG-label
..GCCACCTACTGAACTTCGATACGCCATAATGTCGGATAACCC.
```

20 When this nucleic acid is hybridized to an appropriate probe (an immobilized hexamer in this example) and the probe and ligator-detector are ligated the following structure is obtained

```
    support
              ‛TGAAGCTATGCGGTATTACAGCCTATTGGG-label
..GCCACCTACTGAACTTCGATACGCCATAATGTCGGATAACCC.
```

25

The sequence of the immobilized probe is identified by the location in a probe array where the label of the ligator-detector is detected. The sequence of the adjacent sticky end is identified by the detector array in which the label of the ligator-detector is detected (since a different detector array is used for each

30 sticky end sequence). Finally, the sequence of the recognition sequence is identified by the relationship of the cleavage site to the recognition sequence.

50

Thus, in this example, detection of label in the TGAAGC hexamer position of
the ATAC sticky end detector array indicates the presence of a nucleic acid
fragment in the nucleic acid sample having the sequence
CCTACNNNACTTCGATAC.

5    Relating this sequence to the generalized structure A-B-C-D, A is CCTAC, B is
NNN, C is ACTTCG, and D is ATAC.

When multiple restriction enzymes are used for the first digestion, the
sequence information obtainable can be illustrated with the following structures:
`DNA fragment:`

10   `..NNXXXXNN..NNRRRRNN..NNIIIIIISSSSSSNN..`
`Sequence:        XXXX......RRRR......IIIIIISSSSSS`

In these structures, each character represents a nucleotide. N represents
any nucleotide (having no special identity or relationship to the method). S
represents a nucleotide in the recognition sequence (including sticky end) of the

15   first restriction enzyme. X represents a nucleotide in the recognition/cleavage
site of the second restriction enzyme. R represents a nucleotide in the
recognition sequence of the third restriction enzyme. I represents a nucleotide
complementary to the detector probe. The sequence and distance between the
recognition sites of the second and third restriction enzymes and between the

20   recognition site of the second restriction enzyme and the probe complement are
not determined in the basic method.

From the DNA fragment ...NNXXXX...NNRRRRNN...NNIIIIII
SSSSSSNN..., the sequence information XXXX...RRRR...IIIIIISSSSSS can
be obtained. In this example, the immobilized probes contain hexamer

25   sequences. Each array location where a signal is generated in this example thus
represents a specific sequence : nnnn...nnnn...nnnnnnnnnnnn (where n
represents an identified nucleotide and each ... represents an unidentified gap
sequence). This is referred to as a determined sequence. The portion of the
nucleic acid fragments for which the sequence is determined corresponds to the

30   recognition sequence of the first restriction enzyme, the sequence adjacent to
the recognition sequence to which the detector probe hybridized, the recognition
sequence of the second restriction enzyme, and the recognition sequence of the
third restriction enzyme (S, I, X, and R, respectively).

51

This sequence information can also be represented by the structure

E-B-F-B-C-D

where B is a gap of unknown sequence, C is the sequence to which the probe hybridized, D is the recognition sequence of the first restriction enzyme,

5      E is the recognition sequence of the second restriction enzyme, and F is the recognition sequence of the third restriction enzyme. The gaps represent nucleotides between the recognition sequences of the second and third restriction enzymes and between the recognition sequence of the third restriction enzyme and the sequence to which the probe hybridized. C is always

10     adjacent to the recognition sequence D. In the example above, C is IIIIII, D is SSSSSS, E is XXXX, and F is RRRR.

The sequence information that can be obtained with the disclosed method can be further illustrated using a specific example of a nucleic acid fragment. Assume a nucleic acid sample containing a nucleic acid fragment

15     with the sequence

```
..CGCATGGG..ATAGCTTG..CAAGCTATGGATCCGA..
..GCGTACCC..TATCGAAC..GTTCGATACCTAGGCT..
```

If the sample is first digested with BamHI--a restriction enzyme with a recognition sequence of GGATCC generating a four-base sticky end--the

20     fragment will be cleaved to generate the following fragments:

```
..CGCATGGG..ATAGCTTG..CAAGCTATG        GATCCGA..
..GCGTACCC..TATCGAAC..GTTCGATACCTAG        GCT..
```

When the corresponding adaptor-indexer is ligated to the fragment and the fragment digested with NlaI (recognition sequence CATG) the result is:

25     ```
..CGCATG        GG..ATAGCTTG..CAAGCTATGGATCTGGTATTACAGCCTATTG
..GC        GTACCC..TATCGAAC..GTTCGATACCTAGACCATAATGTCGGATAAC
```
After addition of the second adaptor and hybridization to the corresponding ligator-detector, the following nucleic acid is obtained (sequence from the adaptor-indexer is underlined). Note that the adaptor-indexer hybridizes to both

30     the sticky end sequence and the remaining recognition sequence (that is, the C not in the sticky end).

```
                              GGATCTGGTATTACAGCCTATT-*
CGGTACCTAGAGAGTGTACCC.TATCGAAC..GTTCGATACCTAGACCATAATGTCGGATAA
```

52

When this nucleic acid is hybridized to an appropriate probe (an immobilized hexamer in this example) and the immobilized probe and ligator-detector are ligated the following structure is obtained

```
                                        support
5                                `AGCTATGGATCCGGTATTACAGCCTATT-*
  CGGTACCTAGAGAGTGTACCC.TATCGAAC..GTTCGATACCTAGGCCATAATGTCGGATAA
```

The sequence of the immobilized probe is identified by the location in the array where the label of the ligator-detector is detected. The sequence of the adjacent recognition sequence (including the sticky end) is identified by the

10      array in which the label of the ligator-detector is detected (since a different set of arrays is used for each index sample). The sequence of the recognition sequence of the second restriction enzyme is identified by the array in which the label of the ligator-detector is detected (since a different set of arrays is used for each secondary index sample). Finally, the presence of an internal sequence

15      (the recognition sequence of the third restriction enzyme) is determined by seeing if the signal is absent from the array for the restricted secondary index sample that was digested with the third restriction enzyme (a different array is used for each restricted and non-restricted secondary index sample). If the signal is absent, it indicates the recognition site is present in the fragment. This

20      digestion would be performed prior to the NlaI digestion or following the NlaI digestion if there was an amplification step after digestion with the third restriction enzyme.

Thus, in this example, detection of label in the AGCTAT hexamer position of the TCGA third recognition site array in the GTAC second

25      recognition site set of arrays in the CCTAGG sticky end set of arrays indicates the presence of a nucleic acid fragment in the nucleic acid sample having the sequence

**GTAC...TCGA...TCGATACCTAGG.**

Relating this sequence to the generalized structure E-B-F-B-C-D, C is

30      TCGATA, D is CCTAGG, E is GTAC, and F is TCGA.

Use of labels and sorting in FAAST can be illustrated with the following example which makes use of microbeads in a new way. In this example, 256 ligator-detectors are coupled to the surface of 256 of color encoded microbeads.

53

Each ligator-detector is identified by a single color. After annealing and ligating these new "microbead ligator-detectors" to the target sequences, the 256 sets of "microbead ligator-detectors" are loaded simultaneously in a 4096-well microtiter plate containing 4096 hexamer probes, one probe per well. Only

5      a perfect match during hybridization will ligate a fluorochrome-labeled hexamer to a specific microbead labeled complex. Since the Luminex flow analyzer can distinguish color encoded microbeads and measure their fluorescence simultaneously, one can identify the six bases adjacent to the ligator-detector by knowing the specific address (and the hexamer contained at that address) of the

10     wells on the microtiter plate. The design is illustrated below.

The steps up to generating the single-stranded amplicon are the same as in one basic FAAST, including II-S enzyme cuts, adaptor indexer ligation, second 4 bp enzyme cut, second adaptor ligation, PCR amplification, and capturing and denaturing the fragments to generate single-stranded fragments.

15     The 256 single stranded DNA amplicons are annealed to 256 ligator-detectors. There are 256 different sequences of ligator-detector oligonucleotides, which comprise the 256 different ligator-detectors complementary to each of the 4-base sticky ends generated. The ligator-detector is labeled with a signaling moiety, such as a fluorescent dye or a

20     fluorescent bead. Each of the 256 annealed single stranded preparations is split into 4096 aliquots and hybridized to one of 4096 hexamer probes. For example, the 4096 hexamers are contained in a 4096 microtiter plate, with each well containing a single hexamer probe. Each of the 4096 hexamer probes is fluorescently labelled at the 5'-end and contains a free 3'-hydroxyl end.

25     Following hybridization and ligation of the hexamer probe with the annealed single-stranded amplicons, the 256 preparations are transferred to another well containing streptavidin and unbound material is washed away (the hexamer probe also contains a biotin group). Two signals are then measured.

One signal corresponds to the sample (either tester or control, for

30     instance). Tester and control samples are distinguished by fluorescently encoded beads (Luminex). In one embodiment, one would use 512 colors, 256 colors for the tester and 256 colors for the control. However, one could also use

54

256 color encoded beads where the beads are "offset" between the tester and control. For instance, color 1 corresponds to hexamer 2 for the tester but hexamer 3 for the control, and so on. The second signal derives from the labeled hexamer probe and measures the level of the single-stranded DNA

5     annealed to the ligator-detector. The two signals measured here could be read simultaneously on an instrument such as the Luminex100.

A number of varied probe sets and arrays are known in the art and can be used with FAAST. Terstappen et al. (Tibbe, A.G., *et al., Optical tracking and detection of immunomagnetically selected and aligned cells.* Nat

10    Biotechnol, 1999. 17(12): p. 1210-3; Dolan, G.J. and L.W.M.M. Terstappen, *Magnetic separation apparatus and methods employing an internal magnetic capture gradient and an external transport force,* . 1999: United States Patent 5,985,153; Terstappen, L.W.M.M. and P.A. Liberti, *Quantitative cell analysis methods employing magnetic separation,* . 1999: United States Patent

15    5,993,665; Terstappen, L.W.M.M. and P.A. Liberti, *Methods for biological substance analysis employing internal magnetic gradients separation and an externally-applied transport force,* . 2000: United States Patent 6,013,188) have demonstrated immunomagnetically selected and fluorescently labeled probes for detection of cells of interest. In their technique the cells are labeled using

20    immunospecific binding probes and the resulting labeled cells are induced to move into detector range by an externally applied magnetic field. Lithographic processing of one wall of the sample vessel leads to the improvement of locating the tagged cells along well determined spatial patterns.

Thorp et al. (Napier, M.E., *et al., Probing biomolecule recognition with*

25    *electron transfer: electrochemical sensors for DNA hybridization.* Bioconjug Chem, 1997. 8(6): p. 906-13; Thorp, H.H., C.R. Loomis, and M.E. Napier, *Polymer-electrodes for detecting nucleic acid hybridization and method of use thereof,* . 1999: United States Patent 5,968,745; Thorp, H.H., *et al., Electrochemical detection of nucleic acid hybridization,* . 1999: United States*

30    Patent 5,871,918; Welch, T.W., *Electochemical probes for detection of molecular interactions and drug discovery,* . 1999: WO9964847) have

55

developed an electrochemical detection method. Their method directly detects target nucleic acids without sample amplification or use of fluorescent labels. Detection is accomplished by following the oxidation-reduction cycle of $Ru(bpy)_3^{2+}$ by electrochemical measurement. The measurement elements have

5   synthetically prepared oligonucleotide probes attached to electrodes, the electrodes have been fabricated into a variety of formats including biochips, microtiter plates and hand-held devices.

Spectral labeling coupled with a counting technique, such as flow cytometry, has been exploited for analysis of DNA samples (Fulton, R.J.,

10   *Method and compositions for flow cytometric determination of DNA sequences*, . 1998: United States Patent 5,736,330; Phipps, S., *et al., Precision fluorescently dyed particles and methods of making and using same*, . 1999: WO9919515; Chandler, D. and M.B. Chandler, *Microparticles with multiple fluorescent signals*, . 1999: WO9937814). In such analyses, micron scale beads are color

15   coded according to the probes attached to their surface, targets are labeled with an analytical fluorescent tag, and the coincidence events containing the bead color and the analytical color are counted. In this manner a probe array of many colors can be read out very quickly and easily.

In another technique utilizing microspheres, Walt et al. (Walt, D.R.,

20   *Techview: molecular biology. Bead-based fiber-optic arrays*. Science, 2000. **287**(5452): p. 451-2; Ferguson, J.F., B.G. Healey, and D.R. Walt, *Fiber optic biosensor for selectively detecting oligonucleotide species in a mixed fluid sample*, . 1988: WO9850782; Walt, D.R. and K.L. Michael, *Fiber optic sensor with encoded microspheres*, . 2000: United States Patent 6,023,540; Michael,

25   K.L., *et al., Randomly ordered addressable high-density optical sensor arrays*. Anal Chem, 1998. 70(7): p. 1242-8) have developed a system where the probes are attached to the microspheres and the microspheres subsequently self assemble in a random spatial pattern into the distal end of a fiber optic array. The "optical bar code" signature each microsphere provides the identity of the

30   attached probe, and signal of the labeled target indicates the concentration of the target.

56

**Modification Assisted Analysis of Sequence Tags (MAAST)**

Modification assisted analysis of sequence tags (MAAST) is a form of FAAST that assesses modification of sequences in nucleic acid molecules by detecting differential cleavage based on the presence or absence of modification in the molecules. For example, a site that is methylated in a nucleic acid molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is insensitive to methylation will cleave at that site, thus producing a different pattern of sequence tags. Comparison of the results with different samples of nucleic acids can establish differences in the modification levels or patterns on the different samples. As used herein, a nucleic acid cleaving reagent or restriction enzyme that is sensitive to modification in its recognition site is a nucleic acid cleaving reagent or restriction enzyme that will either cleave only when the site is unmodified or will cleave only when the site is modified (that is, the nucleic acid cleaving reagent or restriction enzyme requires a particular modification state for cleavage). A nucleic acid cleaving reagent or restriction enzyme that is insensitive to modification in its recognition site is a nucleic acid cleaving reagent or restriction enzyme that will cleave regardless of whether the site is modified or unmodified.

MAAST is useful for assessing the state, level, and condition of modifications in nucleic acid molecules. Many nucleic acid modifications are known and most can have biological effects and significance. Methylation, for example, is a universal mechanism for regulating gene expression in animals and plants (Bird and Wolffe, *Methylation-induced repression-belts, braces, and chromatin*, Cell 99:451-454 (1999); Finnegan *et al.*, *DNA Methylation in Plants*, Annual Rev Physiol 49:223-247 (1998); Bird, *DNA Methylation de Novo*, Science **286**:2287-2288 (1999)). Understanding the physiological consequences of methylation has utility in a number of fields. It is well documented that methylation of promoter regions can repress transcription, both *in vitro* and *in vivo* (Baylin *et al.*, *Alterations in DNA methylation: a fundamental aspect of neoplasia*, Adv Cancer Res 72:141-96 (1998)). For instance, the promoters of several genes implicated in neoplasia and tumor suppression are subject to

57

hypermethylation (Melki et al., *Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia*, Cancer Res 59(15):3730-40 (1999)). Methylation also performs important functions in plant development and flowering.

5        The use of differential methylation to study gene function traditionally required prior knowledge of DNA sequences subject to methylation, obtained only after substantial effort to clone, sequence and verify the methylation sensitivity of the region of interest. MAAST expedites the identification of differentially methylated sequences by, in its preferred forms, combining the

10      power of high throughput microarray technology and *in silico* analysis with the sensitivity and quantitation of differential display.

        MAAST has several advantages over other methods currently used to identify promoters on a genomic scale. Current approaches to identify promoter and other regulatory elements in a high throughput manner include: *in silico*

15      analysis of nucleotide sequence for transcription factor binding sites, β-lactamase insertion (Whitney *et al.*, *A genome-wide functional assay of signal transduction in living mammalian cells*, Nat Biotechnol 16(13):1329-33 (1998)), COBRA (Xiong and Laird, *COBRA: a sensitive and quantitative DNA methylation assay*, Nuc Acid Res 25(12):2532-2534 (1997)), and restriction

20      landmark genomic scanning (Costell *et al.*, *aberrant CpG-island methylation has non-random and tumour-type-specific patterns*, Nature Genetics 25:132-138 (2000)).

        Other forms of modification are indicative of certain types of DNA damage caused by particular agencies. These include alkylation, dimerization,

25      derivatization, depurination, or ADP-ribosylation. Examples of modifications and their source are described in Lodovici et al., Levels of 8-hydroxydeoxyguanosine as a marker of DNA damage in human leukocytes, Free Radic Biol Med 28(1):13-7 (2000); Maehira et al., Alterations of protein kinase C, 8-hydroxydeoxyguanosine, and K-ras oncogene in rat lungs exposed

30      to passive smoking, Clin Chim Acta 289(1-2):133-44 (1999); Gamboa Da Costa et al., Characterization of the Major DNA Adduct Formed by alpha-Hydroxy-N-desmethyltamoxifen in Vitro and in Vivo, Chem Res Toxicol 13(3):200-207

58

(2000); Phillips et al., Methods of DNA adduct determination and their application to testing compounds for genotoxicity, Environ Mol Mutagen 35(3):222-233 (2000); Airoldi et al., Carcinogen-DNA adducts as tools in risk assessment, Adv Exp Med Biol 472:231-40 (1999); Purewal et al., Association

5    between acetylator genotype and 2-amino-1-methyl-6- phenylimidazo[4,5-b]pyridine (PhIP) DNA adduct formation in colon and prostate of inbred Fischer 344 and Wistar Kyoto rats, Cancer Lett 149(1-2):53-60 (2000).

MAAST uses the same basic steps as FAAST and includes cleavage with a second nucleic acid cleaving reagent and coupling of a second adaptor to

10   the nucleic acid fragments.  The difference is that MAAST uses a nucleic acid cleaving reagent that is sensitive to modification of its recognition site.  Thus, the nucleic acid cleaving reagent will either not cleave at a site that is modified or cleave only at a site that is modified.  In nucleic acid fragments that have a modification at the recognition site, the fragment will not be cleaved by the

15   second nucleic acid cleavage reagent and no second adaptor will be added to the end.

Uncleaved nucleic acid fragments can be culled from the procedure in a number of ways.  For example, if the nucleic acid fragments are to be amplified following addition of the second adaptors, amplification can be made dependent

20   on the presence of a second adaptor on the end.  This can be accomplished, for example by using a PCR primer complementary to sequence in the second adaptor.  The uncleaved nucleic acid fragments can also be culled by, for example, including a capture tag, sorting tag, or label in the second adaptor.  By capturing or sorting the fragments based on the presence or absence of the

25   capture or sorting tag, one can carry forward in the procedure only those fragments containing a second adaptor.

If a label is associated with the second adaptor, all of the fragments can be carried forward in the procedure but only those having a second adaptor (and thus a label) will produce a detectable signal (in this scenario, the label on the

30   second adaptor would have to be detected in association with a detector probe and/or a ligator-detector).  Use of a label in this manner in the second adaptor can also allow discrimination of cleaved and uncleaved fragments (that is,

59

fragments cleaved or not cleaved by the second nucleic acid cleaving reagent). This can be accomplished by detecting coupling of ligator-detectors to detector probes via labels, capture tags, or sorting tags on the detector probes, ligator-detectors, or adaptor-indexers and also detecting the presence, amount, presence

5     and amount, or absence of a second adaptor on the fragment via its label.

MAAST can also be used to determine how one type of cell or organism influences gene expression or other biological pathways in another type of cell or organism. For example, suppose that a mouse is genetically altered via gene targeting to inactivate a particular methyltransferase gene (there are several known

10    for mouse but for the illustration purposes, assume there is only one). Cells of interest (for instance, B cells) are taken from the mouse and mixed with other cells of interest (for instance, T cells) obtained from a normal mouse. The B and T cells are mixed together. One could then compare the T cell methylation pattern between B cell mixed and non-B cell mixed. It would be safe to conclude that the observed

15    changes in methylation occurred within the T cells since the B cells harbored an inactivated methyltransferase.

Using standard differential gene expression techniques, it would be very difficult for many genes, and impossible for others, to distinguish which transcripts originated in the B cells and which transcripts originated in the T

20    cells (since any two cell types will share expression of a number of genes). However, due to the prior genetic alterations, that is, deleting the methyltransferase gene, MAAST allows examination of gene expression in a mixed cell population.

MAAST can be further understood through the following illustration.

25        1. Isolate genomic DNA (gDNA) from control and tester samples.

2. Select a set of two restriction enzymes. Each set contains a type II-S restriction enzyme and a type II restriction enzyme that will not cut DNA when methylated. In this illustration FokI (II-S) and HinPII (II, methylation sensitive) are used as examples. There a number of type II-S and type II restriction

30    enzymes that could be used.

60

3. First digest gDNA with the type II-S restriction endonuclease. A 5' four base overhang is preferred:

```
5       5'-NNNNNNNNNNNGGATGNNNNNNNNNNNNNNNNN-3'
          |||||||||||||||||||||||||||||||||
        3'-NNNNNNNNNNNCCTACNNNNNNNNNNNNNNNNN-5'
```

Cut with FokI to produce:

```
10      5'-NNNNNNNNNNNGGATGNNNNNNNNN-3'
          |||||||||||||||||||||||||
        3'-NNNNNNNNNNNCCTACNNNNNNNNNNNNN-5'
```

4. Distribute to 256 wells, each with one of 256 unique adapter-

15  indexers.

5. Ligate an adapter-ligator (one of 256) complimentary to the four base overhang. The adapter-indexer contains a biotin moiety to facilitate subsequent purification away from non-ligated strands.

```
20      5'-NNNNNNNNNNNGGATGNNNNNNNNNNnnnn-adap-ind-3'
          ||||||||||||||||||||||||||||||||
        3'-NNNNNNNNNNNCCTACNNNNNNNNNNNNN-adap-ind-5'-Biotin
```

6. Wash ligation products to eliminate non-ligated material. Leaving

25  only ligated products.

```
        5'-NNNNNNNNNNNGGATGNNNNNNNNNNnnnn-adap-ind-3'
          ||||||||||||||||||||||||||||||||
        3'-NNNNNNNNNNNCCTACNNNNNNNNNNNNN-adap-ind-5'-Biotin
```

30      7. Second digest with a methylation sensitive restriction endonuclease (or pair/combination of restriction enzymes that recognize the same site and where one of the enzymes will not cut a methylated site and the other one will).

Example: Control sample digested with HinPII:

```
        5'-NNNGCGCNNNNNNNNNN//NNNNNNNNNNGGATGnnnn..adap-ind-3'
35        |||||||||||||||||   |||||||||
        3'-NNNCGCGNNNNNNNNNN//NNNNNNNNNNCCTACNNNN..adap-ind-5'-Biotin
```

Cut with HinPII to produce:

61

```
5'-NNNG     CGCNNNNNNNNN//NNNNNNNNNGGATGnnnn..adap-ind-3'
   | | | |   | | | | | | | | | |  | | | | | | | | |
3'-NNNCGC    GNNNNNNNNN//NNNNNNNNNCCTACNNNN..adap-ind-5'-Biotin
```

5          Example:  Tester sample (which has undergone methylation at the

HinPII site) digested with HinPII:

```
           CH3
            |
       5'-NNNGCGCNNNNNNNNN//NNNNNNNNNGGATGnnnn..adap-ind-3'
10     | | | | | | | | | | | | | | |  | | | | | | | | |
       3'-NNNCGCGNNNNNNNNN//NNNNNNNNNCCTACNNNN..adap-ind-5'-Biotin
```

Cut with HinPII (methylation sensitive) to produce:

```
15         CH3    (No cut)
            |
       5'-NNNGCGCNNNNNNNNN//NNNNNNNNNGGATGnnnn..adap-ind-3'
           | | | | | | | | | | | | | | |  | | | | | | | | |
       3'-NNNCGCGNNNNNNNNN//NNNNNNNNNCCTACNNNN..adap-ind-5'-Biotin
20
```

8. Wash and ligate on methyl-adapter. The second adapter is specific
for the sticky end generated at the site recognized by the methylation sensitive
restriction endonuclease. Ligation of the second adaptor is most efficient when
the DNA has been cut by the methylation sensitive restriction enzyme (when

25     methylation is not present in this example). When the DNA is not cut, there is
no sticky end available for ligation. Therefore, regions of genomic DNA
methylated in a manner that inhibit cleavage will not be substrates for
subsequent steps.

Example: Ligation of second adapter to DNA cut by methylation

30     sensitive restriction enzyme.

```
5'-sec-adap..NNNG CGCNNNNNNN//NNNNNGGATGnnnn..adap-ind-3'
             | | | |   | | | | | | | |  | | | | | | | | | | | | |
3'-sec-adap..NNNCGC GNNNNNNN//NNNNNCCTACNNNN..adap-ind-5'-Biotin
35

5'-sec-adap..NNNGCGCNNNNNNN//NNNNNGGATGnnnn..adap-ind-3'
             | | | | | | | | | | | | |  | | | | | | | | | | | | |
3'-sec-adap..NNNCGCGNNNNNNN//NNNNNCCTACNNNN..adap-ind-5'-Biotin
40
```

62

9. Amplification of the immobilized gDNA fragment library via PCR or other suitable amplification technologies. Amplification primers are designed to be complementary to the appropriate strand of each of the adapters.

Example: Amplicons generated by PCR

5

```
5'-sec-adap..NNNGCGCNNNNNNNN//NNNNNGGATGnnnn..adap-ind-3'
              | | | | | | | | | | | | | | |    | | | | | | | | | | | | | |
3'-sec-adap..NNNCGCGNNNNNNNN//NNNNNCCTACNNNN..adap-ind-5'
```

10      10. A digestion step with T7 gene 6 exonuclease may be employed to generate single-stranded amplicons. This step requires that one of the primers used for amplification has been synthesized with 5 phosphorothioate linkages substituting for phosphodiester linkages at the 5' end of the primer. The strand containing the phosphorotioate linkages is protected from digestion.

15      Example: Single-stranded amplicon after exonuclease digestion.


```
3'-sec-adap..NNNCGCGNNNNNNNN//NNNNNCCTACNNNN..adap-ind-5'
```


11. Hybridize to ligator-detector probes.

20      Example: gDNA amplicon hybridized with ligator-detector which has a 5' phosphate group. The ligator-detector has a label, such as a fluorescent molecule attached at the 3' end.

```
                                    5'- nnnn-lig-detect-Signal.A
                                        | | | | | | | | | | | | |
3'-sec-adap..NNCGCGNN//NNCCTACNNNNNNNNNNNnnnn..adap-ind-5'
```

25

12. Hybridize to universal detector array. Each of the 256 preparations of genomic DNA containing a hybridized ligator-detector probe is contacted

30      with a comprehensive array of detector probes, consisting of all 4,096 possible hexamers. The hexamer probes on the array are distributed in a particular order. The probes are preferably spaced away from the surface of the array, preferably by a non-polynucleotide linker spacer.

63

Example: Detector-ligator joined to immobilized hexamer.

```
Surface---linker spacer----NNNNNN-3'
```

5
```
                                           5'-nnnn-lig-detect-Signal.A
                                           |||||||||||||||
3'-sec-adap..NNCGCGNN//NNCCTACNNNNNNNNNNnnnn..adap-ind-5'
```

The fluorescent signals in each of the 256 generic detector arrays are

10   measured using a suitable fluorescence scanning instrument or fluorescent

beads.

Example: Tag sequence (5 known + 3 unknown + 10 known bases) for

each array address. The sequence of the methylation site, at some unknown

distance, is also identified.

15
```
        << methyl site     5    (3)   6 + 4
3'-sec-adap..NNCGCGNN//NCCTACNNNNNNNNNNnnnn..adap-ind-5'
```

13. Characterization of derived sequence information. If genomic

20   sequence information available, a BLAST search using the derived sequence

may indicate the presence of a promoter if the sequence lies 5' of a

characterized gene. *In silico* analysis of predicted methylation sites should

facilitate such assumptions. If genomic sequence is unavailable, the sequence

tag can serve as a starting point to facilitate characterization of previously

25   undefined genomic DNA regions of interest.

**Variable Address Analysis of Sequence Tags (VAAST)**

Variable address analysis of sequence tags (VAAST) is a form of the

FAAST method that allows determination of associations, in a nucleic acid

molecule, of different combinations of known or potential sequences. For

30   example, particular combinations of joining and variable regions in

immunoglobulins or T cell receptors can be determined. VAAST uses the same

basic steps as FAAST and adds a step prior to cleavage of the nucleic acid

sample. In VAAST, a recognition site for cleavage, preferably a restriction

enzyme recognition site, is introduced into nucleic acid fragments in the nucleic

35   acid sample. This recognition site is then used as the target of cleavage in the

64

basic FAAST method. The adaptor-indexers should be chosen to match known or potential sequences that would appear adjacent to the sequence into which the recognition site was introduced. The result is fragments with defined end sequences surrounding a central sequence derived from a nucleic acid fragment.

5      This allows the association of known or potential sequences to be assessed. In particular, the association of the sequence into which the recognition site was introduced with a particular adaptor-indexer (which has sequence matching the known or potential adjacent sequence) can be detected.

VAAST can be further understood through the following illustration.

10      1. Isolate RNA and make cDNA (DNA) from the samples of interest.

2. Select a set of two PCR primers, one specific for the V region, one specific for the J region (or D or C region if desired). In the first preferred method (there are several), the V primer contains a Type II-S restriction site, say FokI. Furthermore, the V specific primer contains several phosphorothioate

15      linkages at the 5' end. The length of the amplicon is controlled by placement of the primers.

Example: cDNA target:

```
          V            | (N) |      J      /  /     C
5'-GATCTAGATCAGTNNNNNNTGACCATGT....CGTGAATCATG-3'
  | | | | | | | | | | | | | | | | | | | | | | | | | | | |     | | | | | | | | | |
3'-CTAGATCTAGTCANNNNNNACTGGTACA....GCACTTAGTAC-5'
```

Example of cDNA amplicon, with introduced FokI site:

```
         V             | (N) |      J
5'-GATCTGGATGAGTNNNNNNTGACCATGT....-3'
  | | | | | | | | | | | | | | | | | | | | | | | | | | | |
3'-CTAGACCTACTCANNNNNNACTGGTACA....-5'
```

3. After PCR, the amplicons are digested with FokI and an adapter-
30      indexer is ligated. This readies the amplicon for digestion with T7 exonuclease gene 6. The 5' end of the adapter-indexer is not protected from T7 exonuclease

65

gene 6 digestion.   Hence, a digestion step with T7 exonuclease gene 6 will

generate single-stranded amplicons.

cDNA amplicon from step 2


5                    FokI digestion

```
              V        | (N) |           adapter-indexer
    5'-GATCTGGATGAGTNNNNNN-3'      5'-TGACXXXXXXXXX-3'
      | | | | | | | | | | | | | | | | | | |          | | | | | | | |
    3'-CTAGACCTACTCANNNNNNACTG-5'      3'-YYYYYYYYY-5'
```
10

   4.  After PCR, the amplicons are digested with T7 exonuclease gene 6.

   Example of T7 exonuclease gene 6 treated amplicon:

```
         V_x      |  (N)  |   J_y  | adapter-indexer
    5'-GATCTGGATGAGTNNNNNNTGACXXXXXXXXXX.....-3'
```
15

   5.  Single stranded amplicons are hybridized to a pool of 100 V ligator-

detectors, each containing a label and each specific for a single V region gene

segment.  All V ligator-detectors (about 25 bases long) are designed a priori to

hybridize specifically to a single V region gene segment.  The V and J subscript

20    numbers are chosen arbitrarily here for illustration purposes.


```
            V_1            |  (N)  |      J_10
        5'-GATCTAGATCAGTNNNNNNTGACCATGT.....-3'
    3'-signal-ligator-detector-V_1-5'
```
25

```
            V_1            |  (N)  |      J_1
        5'-GATCTAGATCAGTNNNNNNACTGATCGG.....-3'
    3'-signal-ligator-detector-V_1-5'
```

```
            V_2            |  (N)  |      J_10
30      5'-GTTATACTGCAACNNNNNNTGACCATGT.....-3'
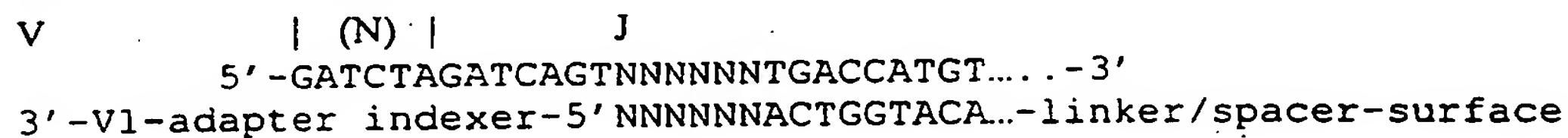    3'-signal-ligator-detector-V_2-5'
```

   Note that in this example, both $V_1$ ligator-detector and $V_2$ ligator-

35    detector will generate a signal by hybridizing and ligating to the $J_{10}$ probe.  In

addition, $V_1$ ligator-detector will generate a signal from the $J_1$ address; hence the

name variable address analysis of sequence tags.  The address of the hybridized


66

complex will be dependent upon which J region gene segment the V gene segment has joined to. This same principle would hold true for all 100 V gene segments.

6. The amplicon/adapter indexer complexes are contacted with a comprehensive address array containing oligonucleotide probes specific for all known J gene segments and those sequences created by the addition of 6 nucleotide bases between the V and the J. Because of the small number of J segments (50), it is not necessary to limit the size of the probes to 6 bases. A preferred length may be 8 to 15 bases or longer if more stringent conditions are desired to increase hybridization fidelity.

```
V              |  (N) |          J
        5'-GATCTAGATCAGTNNNNNNTGACCATGT.....-3'
        3'-V1-adapter indexer-5'NNNNNNACTGGTACA..-linker/spacer-surface
```

Illustration of probe complexity required to cover all possible sequences generated at the N region:

| Number of N region bases | Number of possible N region sequences | Number of Js | Number of sequences |
|---|---|---|---|
| 0 | 0 | 50 | 50 |
| 1 | 4 | 50 | 200 |
| 2 | 16 | 50 | 800 |
| 3 | 64 | 50 | 3,200 |
| 4 | 256 | 50 | 12,800 |
| 5 | 1024 | 50 | 51,200 |
| 6 | 4096 | 50 | 204,800 |
|   | 5461 |   | 273,050 |

Therefore, assuming all possible sequences are generated at the N region and that all J gene segments are utilized, there would be 273,050 sequences possible. If one further assumes that all 100 V gene segments are used, then the total number of sequences possible would be:

100 X 273,050 = 27,305,000 - just over twenty seven million sequences.

The following observations expand the sequence complexity that could be interrogated 10-100 fold:

Observation 1: Although not taken into account here, it is well established that N region additions can exceed six bases, although zero to six

5   appear to be the most common.

Observation 2: Additional complexity comes from a phenomenon known as DNA editing, in which bases at the 3' end of the V region and the 5' end of the J region are "nibbled" back to generate still more diversity. For simplicity, DNA editing is not accounted for in this illustration.

10   Observation 3: Finally, the Ig and TCR receptor families described here are expressed as heterodimers on the cell surface. For instance, the TCR-alpha chain associates with the TCR-beta chain. It is therefore a simple task to generate a chip that could simultaneously interrogate a cDNA preparation for sequence variability in both the alpha and beta genes.

15   Observation 4: The full diversity of the TCR or Ig sequences assumes any combination of V, D, J, and C regions. The potential combinations include VDJC, VNDJC, VJC, VNJC, VDDJC, VNDDNJC, VDNDNJC, VNDNDNJC, and any combination of these.

**Mass Spectroscopy Detection**

20   Mass spectrometry techniques can be utilized for detection in FAAST. These techniques include matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy. Such techniques allow automation and rapid throughput of multiple samples and assays.

Mass spectrometry detection works better with smaller molecules so it is

25   useful to cut some FAAST components prior to, or as part of mass spectrometry detection. A number of methods are contemplated where an oligonucleotide molecule to be detected is cut to a shorter length prior to detection by mass spectrometry. The FAAST protocol would proceed as normal and, in the preferred embodiment, the surface that has the hexamer probes attached would

30   be compatible with the source region of a matrix assisted laser desorption

68

ionization, time of flight, mass spectrometer (MALDI-TOF-MS). The resultant

fragment of the FAAST process with FokI index cut would look something like

```
Surface---3' HHHHHHNNNIIIXXXXXXXXXXXXXXXXXXXXXXXX-L-3'
             XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-5
            /
3'...YYYYYYYY
```

5

Where:

H are the hexamer probe, ligated to the fragment;

10       I are the quadramer from the indexing step;

N are the unknown bases not read out by FAAST;

X are complementary bases, not germane;

Y are the remaining nucleotides of the FAAST fragment;

- are linker, universal bases, mimics or other analogs;

15       L is a label;

bold indicate a possible sub-fragment.

For fragments of greater than approximately 50 bases the performance

of mass spectrometry techniques degrades for DNA samples. Chemical,

biological, physical (thermal), and other cleaving reagents can be used to

20       generate smaller, more optimal, sub-fragments to be analyzed in the mass

spectrometer. The degree of fragmentation is somewhat tunable in instruments

like the Q-TOF systems (Micromass, US head office at Suite 407N, 100

Cummings Center, Beverly, MA 01915-6101, USA.) where one can look at the

parent ion, then increase the fragmentation to see the decomposition fragments

25       and thus the sequence; such a technique is contemplated to determine the full

sized sub-fragment, and infer the sequence (which is longer sequence

information than for the basic FAAST method) of the sub-fragment through

these known tools. The detectable fragment can be top strand, bottom strand, or

both strands depending upon the scheme. The label may be a cleavable mass

30       tag or the strand need not be labeled.

There are several useful cleaving reagents for this purpose. For

example, one technique is that of Szybalski (described elsewhere herein) where

FokI is used to cut at a fixed distance from an arbitrary, specific, recognition

site. This technique can be extended to other restriction enzymes of type II-S or

69

type III.   One could also use this technique twice, once to trim off the end

nearer the surface, once to trim off the end further from the surface; preferably

one would use a type II enzyme to cut the end furthest from the surface.

Use of McrBC (NEB, #272), can be used to cut at methylcytosine sites

5       adjacent to G/A.  The cut site is not well defined (approximately 30 bases)

which may be used to advantage to generate the parent as well as the

fragmentation set.  Metal containing porphyrins attached to oligonucleotides

have been shown to cut DNA very near the porphyrin when exposed to light

(texaphyrins, US5607924).  One could denature and use a hybridization

10      texaphryin and light to cleave the remaining strand.  Another cleavage

technology is that of Dervan (Cartwright, I.L., *et al.*, *Cleavage of chromatin*

*with methidiumpropyl-EDTA . iron(II)*. Proc Natl Acad Sci U S A, 1983.

**80**(11): p. 3213-7; Schultz, P.G. and P.B. Dervan, *Sequence-specific double-*

*strand cleavage of DNA by penta-N- methylpyrrolecarboxamide-EDTA X*

15      *Fe(II)*. Proc Natl Acad Sci U S A, 1983. **80**(22): p. 6834-7).  Techniques using

photocleave linkages are described by Olejnik et al. (Olejnik, J., *et al.*,

*Photocleavable peptide-DNA conjugates: synthesis and applications to DNA*

*analysis using MALDI-MS*. Nucleic Acids Res, 1999. **27**(23): p. 4626-31;

Olejnik, J., E. Krzymanska-Olejnik, and K.J. Rothschild, *Photocleavable*

20      *affinity tags for isolation and detection of biomolecules*. Methods Enzymol,

1998. **291**: p. 135-54; Olejnik, J., E. Krzymanska-Olejnik, and K.J. Rothschild,

*Photocleavable aminotag phosphoramidites for 5'-termini DNA/RNA labeling*.

Nucleic Acids Res, 1998. **26**(15): p. 3572-6; Olejnik, J., E. Krzymanska-

Olejnik, and K.J. Rothschild, *Photocleavable aminotag phosphoramidites for*

25      *5'-termini DNA/RNA labeling*. Nucleic Acids Res, 1998. **26**(15): p. 3572-6;

Olejnik, J., *et al.*, *Photocleavable biotin derivatives: a versatile approach for*

*the isolation of biomolecules*. Proc Natl Acad Sci U S A, 1995. **92**(16): p. 7590-

4) These linkages can be cleaved using light to release the fragment from the

surface, thus allowing one to provide a more gentle desorption.  WO 0004036

30      describes photocleavable nucleotides and methods for their use.

In one embodiment, a mass label such as peptide nucleic acid (PNA) molecules (Hanvey et al., Science 258:1481-1485 (1992)) of different sequence and molecular weight can be used as labels that bind specifically to sequence in ligator-detectors or adaptor-indexers. Laser desorption of the samples is used to

5    generate MALDI-TOF mass spectra of the PNA labels, which are released into the spectrometer and resolved by mass. The intensity of each PNA label reveals the relative amount of different components (e.g. ligator-detectors or adaptor-indexers. In other words, the PNA spectra generate scalar values that are indirect indicators of the relative abundance of the labeled component at

10   specific locations in an array.

**Probability Detection**

Sequencing by hybridization is known to produce mismatch errors (Lipshutz, R.J., *Likelihood DNA sequencing by hybridization.* J Biomol Struct Dyn, 1993. 11(3):637-53). Database searching for sequence information

15   currently is regular expression based and requires matched "letters" between the database entry and the search sequence. FAAST allows replacement of regular expression matching (match versus no-match per base) with a probability function to determine a confidence in the assignment of the identity of a sequence tag (that is, the fragments produced in FAAST).

20   The FAAST method uses ligation to improve the specificity of the hybridization near the ligation site. Despite this improvement, there will remain a finite probability of a mismatch, particularly for nucleotides more removed from the ligation site. The error rate depends on least two mismatch properties:

base pairing, i.e. A with G;

25   distance from the ligation site.

As an illustration of the process to determine the confidence value, consider the two bases in a hexamer probe furthest from the ligation site, numbering the bases as shown here.

```
                    <hexamer>
30   surface-linker-spacer-NNNNNNnnnn-ligator-detect-signal-3'<probe>
                    ||||||||||||||||||||||||||||
            3'-adapter..NNNNNNNNNNnnnn-adapter-indexer-5'   <target>
                    123456 <position>
```

71

where for this particular case one has, surface---linker-spacer------ATXXXX, focusing on the AT (positions 1 and 2) bases for purpose of the immediate illustration.

To evaluate the possible set of sequences represented, weight matrices are used, following Dayhoff (Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary changes in proteins*, in *Atlas of Protein Sequence and Structure*, M.O. Dayhoff, Editor. 1978, National Biomedical Research Foundation: Washington DC) and Venezia (Venezia, D. and P.J. O'Hara, *Rapid motif compliance scoring with match weight sets*. Comput Appl Biosci, 1993. 9(1):65-9) protein techniques. The coefficient in these matrices will be determined experimentally for the FAAST system. Below is an example of matrices (with illustrative coefficients) representing position 1 and 2, where the columns represent the upper strand nucleotide and the rows represent the lower strand nucleotide. The actual coefficients can be determined empirically.

```
        Position 1                    Position 2
      A    T    C    G            A    T    C    G
  A[.02,.90,.03,.05]          A[.01,.97,.01,.01]
  T[.90,.02,.03,.05]          T[.97,.01,.01,.01]
  C[.02,.03,.05,.90]          C[.01,.01,.01,.97]
  G[.03,.02,.90,.05]          G[.01,.01,.97,.01]
```

For the case of a perfect match detection on the hexamer ATXXXX the score is determined to be the product of the coefficients of the matrices, shown below here in bold; 0.90 x 0.97 = 0.87.

```
        Position 1                    Position 2
      A    T    C    G            A    T    C    G
  A[.02,.90,.03,.05]          A[.01,.97,.01,.01]
  T[.90,.02,.03,.05]          T[.97,.01,.01,.01]
  C[.02,.03,.05,.90]          C[.01,.01,.01,.97]
  G[.03,.02,.90,.05]          G[.01,.01,.97,.01]
```

A case where a singe base mismatch in one strand occurs, for example A→G in position 1 on the hexamer side, the score is determined in a similar fashion, to be 0.05 x 0.97 = 0.05

```
         Position 1                 Position 2
5         A    T    C    G           A    T    C    G
      A[.02,.90,.03,.05]         A[.01,.97,.01,.01]
      T[.90,.02,.03,.05]         T[.97,.01,.01,.01]
      C[.02,.03,.05,.90]         C[.01,.01,.01,.97]
      G[.03,.02,.90,.05]         G[.01,.01,.97,.01]
```

10      This procedure can be extended to an arbitrary number of bases in a similar manner. For a given number of nucleotides the score can be computed for all possible mismatches and rank ordered to reveal the most probable identity. A cut-off score can be used to reduce the number of possible identities from the matrix estimation. For example using the example matrices above,

15      sequences with a threshold score above 0.50 would yield only one sequence, that being a sequence which matches the probe.

This method of estimating sequences and their respective probability scores from the universe of mismatch events for a said probe can from extended from 1 to n, where n is the number of free bases available for hybridization.

20      In an organism that has not been completely characterized (i.e. at least sequenced and consensus sequence assembled) one can compute a confidence value for uniqueness if one assumes a random distribution of bases. For example, if one has a candidate of 15 bases in length, in an organism which has an estimated $10^8$ base genome, one expects the 15 base fragment to be unique

25      because $10^8/4^{15} = 0.1$ is much less than 1. The genome would have to be 10 times larger before one would expect an occurrence of two instances of the particular 15 base fragment.

The distributions, in known genomes, are known not to be completely random and the initial assumption of a random distribution can be improved as

30      information is gathered. This new information can be used to assign and use confidence values.

73

As an example, consider a fictitious gene family ABCD, whose
members are ABCD1, ABCD2 and ABCD3. The three members were
discovered following some event such as heat shock, and they are thus
putatively assigned to belong to the heat shock family of genes and happen to

5      have significant stretches of conserved sequence among the family of genes.
Also consider the organism to be a plant, where ABCD1 was isolated from the
plant root, ABCD2 was isolated from the plant leaf, and ABCD3 was isolated
from the plant flower. The estimation matrix may look like


10                    1      2      3
ABCD1[.60,.15,.05]
ABCD2[.25,.60,.15]
ABCD3[.05,.15,.60]
where the column 1 represents root, column 2 represents leaf and column 3

15     represents flower.

In a single experiment where one has high confidence in the sequence
but the sequence may belong to one of the three known members of the family,
the source of the sample (i.e. root, leaf or flower) allow estimation of the
identity of the gene. For the fully mathematically closed treatment the matrix

20     must contain all elements of the family, here to allow for a still to be found gene
in this family, the rows and columns do not add to 1; all the other members are
assigned a sum of 0.05, the values to be updated as the amount of information
known about the organism increases.

One can extend this estimation to include organism homology. That is,

25     if one were to search a database of all organisms for a FAAST tag from gene
ABCD1 of Plant 1 there may be matches to Plant 2, Plant 3, Mammal 1, etc.
The estimation matrix would be constructed from the known organism data in
the database.

The calculations and analysis described above can be illustrated using

30     the following example of construction of a catalog. Consider a two probe
detector array, a control sample, and a tester sample. Consider the two probes
to have the known sequences: A, <substrate--linker--AGGGAG-3'>, and, B,

74

<substrate--linker--ATGGAG>. These probes will capture their cognate

sequence: AA, <...TCCCTC...>, and, BB, <...TACCTC...> from the control

and tester samples, as well as some mismatched species with lower probability

as described herein. Utilizing the estimation matrix technique as discussed

5      above one calculates the probabilities of the correct matching.

The FAAST procedure is conducted on the control and tester, resulting

signals are collected from the probe detector array, and a catalog is made which

contains the four signals:

|        | control |     |        | tester |     |
|--------|---------|-----|--------|--------|-----|
|        | AA      | BB  |        | AA     | BB  |
| A      | .30     | .03 | A      | .80    | .10 |
| B      | .03     | .50 | B      | .03    | .50 |

The catalog also contains the probabilities, and/or entries derived from

the probabilities, for each probe/target combination, as discussed above. For

15     purpose of illustration, let us assume that the probability of having probe

sequence A paired with target sequence AA is 0.80, and the probability of

having probe sequence A paired with sequence BB is 0.10, probe sequence B

paired with target sequence AA is 0.05, and the probability of having probe

sequence B paired with sequence BB is 0.75, or

20            estimation

|   | AA  | BB  |
|---|-----|-----|
| A | .80 | .10 |
| B | .05 | .75 |

It is a simple matter of application of linear algebra to determine the

25     signals corresponding to each target. Here, for example, multiplying the

corresponding entries together to convert the control and tester to the pattern

corresponding to the probabilistic pattern of the target of interest. For example,

the total signal ascribed, in the control sample, to AA target is 0.30 x 0.80 (on A

probe site, perfect match) + 0.03 x 0.05 (on B probe site, imperfect match) =

30     approximately 0.24. On the tester sample, the AA target signal is 0.80 x 0.80 +

0.03 x 0.05 = approximately 0.64. Comparison of the pattern for the control

and tester, for the sequence corresponding to AA, exhibits an increase in the

75

relative amount of AA from 0.24 to 0.64 for control to tester respectively. All

other entries in the pattern are calculated in the same fashion.


**Illustrations**

5          The disclosed method can be further understood by reference to the

following illustrations.

### Illustration 1 - FAAST

1. Double-stranded cDNA is prepared with reverse transcriptase, using

standard procedures.

10         2. The double-stranded cDNA is cleaved with a class II-S restriction

endonuclease. Any one of several available class II-S enzymes may be used or

this step. A preferred class II enzyme will recognize a unique 5-base sequence

in the cDNA, and will cleave at a site located 8 or 9 bases downstream of the

recognition sequence. The number of possible cleavage sites comprises a total

15   of 256 different sequences, resulting in the generation of any of 256 different 4-

base sticky ends. Examples of suitable enzymes for this step are BbvI

(recognition, GCAGC -> N8 cleavage), and FokI (recognition, GGATG -> N9

cleavage).

Example: cleavage with FokI

20

```
NNNNNNNNGGATGNNNNNNNNNN-3'
||||||||||||||||||||||||
NNNNNNNNNCCTACNNNNNNNNNNNNnnnn  -5'
```


25         3a. Using 384-well microtiter dishes, the cDNA digest is separated into

256 equal aliquots, and each aliquot is incubated with a first adaptor-indexer in

the presence of T4 DNA ligase. The process of linking adaptors to the sticky

ends generated by class II-s enzymes has been denominated "indexing" (Unrau

and Deugau, *Gene* 145:163-169 (1994)). There are 256 different first adaptor-

30   indexers, corresponding to the 256 possible 4-base sticky ends generated by the

cDNA digest. The adaptors contain a biotin moiety at the non-ligating end, so

as to enable immobilization of the adaptor in a subsequent step.


76

Example:  ligation of first adaptor indexer (one of 256)

```
NNNNNNNNGGATGNNNNNNNNNN          nnnn-adaptor-indexer-3'
|||||||||||||||||||||            |||||||||||||||
5    NNNNNNNNNCCTACNNNNNNNNNNNnnnn       -adaptor-indexer-Biot


NNNNNNNNGGATGNNNNNNNNNNNnnnn-adaptor-indexer-3'
||||||||||||||||||||||||||||||||||||||||||||||||||
10   NNNNNNNNNCCTACNNNNNNNNNNNnnnn-adaptor-indexer-Biot
```

3b.  The ligated material is incubated in the presence of phage T4
endonuclease VII, under mismatch cleavage conditions (Youil et al., *Genomics*
32:431-435 (1996)).  This step will cleave any ligated DNA where the adaptors

15   had been joined to an imperfectly paired sticky end.  Thus, this step performs
error-checking of the previous ligation step.  This step is optional.

4a.  Each of the 256 cDNA aliquots is digested with a single specific
restriction endonuclease that recognizes a unique 4-base sequence.  Any four-
cutter enzyme that is a frequent cutter is suitable for this step.  This digestion

20   will result in cleavage of the majority of those cDNAs that already contain a
ligated first adaptor-indexer.  After this digest, the cDNA fragments will
comprise a heterogeneous population with a mean size of 100 to 200 bases,
depending on the specific pair of class II-S and 4-cutter restriction enzymes
used.  A cDNA fragment will comprise a useful tag for future analysis if it

25   contains a ligated adaptor and a sticky end.  This step is optional.

4b.  The material from each of the 256 reactions is placed on a separate
microtiter well containing immobilized avidin, in order to bind the biotinylated
adapters on the surface.  This step will result in the immobilization of all the
fragments that were successfully ligated to adaptors in step 3.  Any unbound

30   fragments are removed by washing.  The resulting collection of 256
immobilized fragment sets is called the immobilized cDNA fragment library.
This step is optional.

77

Example: Surface-immobilized tag with a four-cutter sticky end

```
5'nnnnNNN...NNNGGATGNNNNNNNNNNnnnn-adaptor-indexer-3'
   ||||||||||||||||||||||||||||||||||||||||||||||
   3'NNN...NNNCCTACNNNNNNNNNNNnnnn-adaptor-indexer-Biot-SURF
```

5

5. Each member of the immobilized cDNA fragment library is incubated
with a second adaptor, in the presence of DNA ligase, in order to join the
immobilized cDNA fragments to a universal priming sequence. Unligated

10     adaptors are removed by washing. This step is optional.

Example: Immobilized cDNA tag with two adaptors

```
second.adap-nnnnNNN...
|||||||||||||||||||||||
second.adap-nnnnNNN...
```

15

```
...NNNGGATGNNNNNNNNNNnnnn-adaptor-indexer-3'
   |||||||||||||||||||||||||||||||||||||||||||||
...NNNCCTACNNNNNNNNNNNnnnn-adaptor-indexer-Biot-SURF
```

20

6a. All members of the immobilized cDNA fragment library are
amplified using a suitable amplification method, such as PCR. A total of 256
separate amplification reactions are performed. Two amplification primers are
used, one designed to be complementary to a strand of the second adaptor, and

25     another designed to be complementary to the first adaptor-indexer. The primer
complementary to the first adaptor-indexer is one of 256 possible primer
sequences, overlapping by four bases with only one of the 4-base sticky ends
generated in step 2. Amplification is performed for a total of 12 to 18 cycles
using PCR. This step is optional.

30          Example: Amplicons generated by PCR

```
second.adap-nnnnNNN...NNNGGATGNNNNNNNNNNnnnn-adaptor-indexer-3'
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
second.adap-nnnnNNN...NNNCCTACNNNNNNNNNNNnnnn-adaptor-indexer-5'
```

35

78

6b.  A digestion step with T7 exonuclease gene 6 (Nikiforov et al., *PCR Methods and Applications* 3:285-291 (1994)) may be employed to generate single-stranded amplicons.  This digestion step assumes that one of the primers used in 6a had been synthesized with 5 phosphorothioate linkages substituting

5    for phosphodiester linkages at the 5' end of the primer (Nikiforov et al. (1994)). These linkages protect one strand from digestion.  This step is optional.

Example:  Single-stranded amplicon after exonuclease digestion

```
second.adaptor-nnnnNNN..NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```
10

6c.  There exist alternative methods for generating amplified tags instep 6.  One example is the Strand Displacement Amplification method of Walker et al., *Nucleic Acids Research* 20:1691-1696 (1992).  Yet another method for amplification of cDNA tags is rolling Circle Amplification (Lizardi et al.,

15   *Nature Genetics* 19:225-232 (1998); U.S. Patent No. 5,854,033 to Lizardi).  In order to use rolling circle amplification, it is necessary to generate circularized cDNA tags by ligation.  Ligation methods for generating circularized double stranded DNA are well known in the art.  This step is optional.

7a.  Each of the 256 preparations of amplified cDNA tags is transferred

20   to another 384-well microtiter dish and hybridized with a fluorescence-labeled detector-ligator oligonucleotide.  There are 256 different sequences of ligator-detector oligonucleotides, which comprise the 256 different ligator-detectors complementary to each of the 4-base sticky ends generated in step 2.  Only the specific cognate sequence is used for each of the 256 amplified cDNA tags.

25   The ligator-detectors contain a 5'-phosphate group.

Example:  cDNA amplicon hybridized with phosphorylated ligator-detector

```
                              5'-nnnn-ligator-detect-
30   SIGNAL.A
                              |||||||||||||||||||||||
     2nd.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

7b. For more accurate expression profiling of control vs. tester mRNA
populations, this step may incorporate a second set of detector-ligator
oligonucleotides labeled with a fluorescent dye of a different color. Thus,
control tags are hybridized with 256 oligonucleotides of color "A", while tester

5      tags are hybridized with 256 oligonucleotides of color "B". After hybridization
is complete, paired sets of cDNA tags are mixed, and one proceeds to step 8
below.

Example: cDNA amplicon hybridized with phosphorylated ligator-
detector

10

```
                                        5'-nnnn-ligator-detect-
SIGNAL.B

                         | | | | | | | | | | | | | | | | | | | |
2nd.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

15

8. Each of the 256 preparations of cDNA, which now contain
hybridized ligator-detectors, is contacted with a comprehensive probe array.
The comprehensive probe array consists of a complete set of 4,096 hexamers
(or, alternatively, 16,384 heptamers), ordered on a solid surface, and separated

20      from the surface by a non-polynucleotide linker-spacer. The hexamers (or
heptamers) contain free 3'-hydroxyl ends. Each of the 256 detection reactions
on the arrays is performed under identical hybridization/ligation conditions,
using T4 DNA ligase or any other suitable ligase, in order to achieve ligation of
a specific, cognate hexamer sequence to the corresponding sequence adjacent to

25      the 5'-end of a detector ligator probe. After hybridization, any unbound
material is removed by very stringent washing.

Example: Detector-ligator joined to immobilized HEXAMER

```
                         Hexamer
30      SURFACE-linker-spacer-NNNNNN-3'

                                        5'-nnnn-ligator-detect-
```

80

```
SIGNAL.A
                                  | | | | | | | | | | | | | | | | | | | |
        2ⁿᵈ.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

5

```
                              Hexamer
            SURFACE-linker-spacer-NNNNNNnnnn-ligator-detect-
SIGNAL.A
                                  | | | | | | | | | | | | | | | | | | | | | | |
        2ⁿᵈ.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

10

```
                              Hexamer
            SURFACE-linker-spacer-NNNNNNnnnn-ligator-detect-
SIGNAL.B
                                  | | | | | | | | | | | | | | | | | | | | | | | |
        2ⁿᵈ.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

15

9. To minimize hybridization errors, an error-checking step is performed after ligation. This error-checking step will cleave incorrectly paired

20   bases (those containing mispaired C or T) in the ligator-detector probe. The step is performed using the chemical cleavage of mismatched bases as described (Ellis et al., *Human Mutation* 11:345-353 (1998)). While this method will not destroy all incorrectly ligated ligator-detector probes, it will reduce noise at approximately 50% of the array addresses. Other error-checking

25   methods could be used to remove mismatched G's and A's. This step is optional.

10. The fluorescent signals in each of the 256 generic probe arrays are measured using a suitable fluorescence scanning instrument. An expression profile dataset is generated from the combined signals of all 256 arrays. The

30   address of each signal is translated into a known sequence, shown in bold letters.

Example: Determined sequence (5 known + 3 unknown + 10 known bases) for each array address.

35

```
                          5   (3)   6   + 4
    second.adap-nnnnNNN...NNNCCTACNNNNNNNNNNnnnn-adaptor-indexer-5'
```

11. The expression profile dataset is further analyzed taking into consideration the expected address signatures of highly abundant gene products, which can be predicted from available sequence data. The expression differences between two tissues are measured by the ratio of the two different

5    color signals corresponding to control and tester. A similar experiment should be performed with a different pair of restriction enzymes, in order to resolve inconsistencies generated by cross-hybridization of abundant sequences at incorrect addresses.

Oversampling calculations:

10   Average mRNA = 17000 bases, average tag = 120 bases

Number of tags for the average mRNA = 4

Thus, the average mRNA is sampled many times, increasing the likelihood of accurate detection.

Considerations on complexity of the sieve:

15   mRNA complexity = 12,000 transcripts*1700 bases ==>>1 X $10^7$

Pentamer=1/1024 * Tetramer=1/256 * Hexamer=1/4096 ==>>> 1/1X $10^9$

Assuming average tag size of 120 bases:

Number of tags = 12000*(4) = 48,000 tags

Number of tags in each of the 256 microtiter wells = 188

20   Number of different array addresses = 256 * 4,096 = 1,048,576  (Hexamer)

Number of different array addresses = 256 * 16,384 = 4,194,304  (Heptamer)

### Illustration 2 - FAAST

1. Double-stranded cDNA is prepared with reverse transcriptase, using standard procedures.

25       2. First Digest:  The double-stranded cDNA is separated into C equal aliquots (index samples), and each index sample is cleaved with a different six-base recognition restriction endonuclease that generates a sticky end of 2 or 4 bases.  Any one of many available restriction endonucleases may be used for this step.

-82-

Examples of suitable enzymes for this step include:

BamHI        GGATCC

BclI         TGATCA

BglII        AGATCT

5    BsiWI        CGTACG

BspHI        TCATGA

EcoRI        GAATTC


Example: cleavage with one of the C first restriction enzymes (R is a

10    base in the recognition sequence, S is a base in the sticky end (which is also in

the recognition sequence, of course), and N is any base):

```
...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNR-3'
...|||||||||||||||||||||||||||||||||
...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNRSSSS -5'
```

15

3. First ligation:  Each of the C index samples of cleaved, double-

stranded cDNA is mixed with a double stranded adaptor-indexer containing the

correct sticky end for the corresponding enzyme used in the first digest, and an

arbitrary primer complement sequence.  There are C different such adaptor-

20    indexers, and each adaptor-indexer has the following structure:

primer complement

```
sticky end  SSSSnnnnnnnnnnnnnnnnnnnnnnn
            |||||||||||||||||||||||
            nnnnnnnnnnnnnnnnnnnnnnn
```

25                                    primer sequence of adaptor-indexer

A suitable DNA ligase is added, and the adaptor-indexer is thus ligated

to all compatible ends present in the cleaved cDNA.

4. Second Digest:  Each of the C index samples, containing ligated,

double-stranded cDNA, is separated into R (typically 1 to 5) equal aliquots

30    (secondary index samples), and each of these secondary index samples is

cleaved with a different mixture of N (typically 1 or 2) different four-base

recognition restriction endonuclease.  Any mixture of the many available

restriction endonucleases may be used for this step, provided that this mixture

does not contain the same enzyme used in the first digest.  Thus there are C * R

83

possible fractions generated in the process of performing the second digest. The
mixture of N enzymes used in the second digestion step can include enzymes
such as the following:

|       | AciI   | sticky = 2 | CCGC |
|-------|--------|------------|------|
| 5     | Hha I  | sticky = 2 | GCGC |
|       | TaqI   | sticky = 2 | TCGA |
|       | Csp6I  | sticky = 2 | GTAC |
|       | Sau3AI | sticky = 4 | GATC |
|       | Tsp509I| sticky = 4 | AATT |
| 10    | TaiI   | sticky = 4 | ACGT |

This second digestion with the mixture of N enzymes will result in
cleavage of many of those cDNAs that already contain a ligated adaptor-
indexer. After this digest, the cDNA fragments will comprise a heterogeneous

15      population with a mean size of 60 to 300 bases, depending on the specific
mixture of N four-base recognition restriction endonucleases used.

5. Each aliquot is incubated with a second adaptor in the presence of T4
DNA ligase. There are N different second adaptors corresponding to the
possible 2-base or 4-base sticky ends generated by the mixture of N enzymes.

20      The second adaptors may contain a biotin moiety at the non-ligating end, so as
to enable immobilization of the second adaptors in a subsequent step.

Example: ligation of second adaptor (one of N)

```
                                      cDNA tag           primer complement

Biot-2nd.adaptor          5'-ssssNNN...NNNRSSSSnnnnnnnnnnnnnnnnnnnnnnnn
25   |||||||||||                     |||...|||||||||||||||||||||||||||||||
     2nd.adaptor-ssss-3'             NNN...NNNRSSSSnnnnnnnnnnnnnnnnnnnnnnnn


                                      cDNA tag           primer complement

Biot-2nd.adaptor-ssssNNN...NNNRSSSSnnnnnnnnnnnnnnnnnnnnnnnn
30   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     2nd.adaptor-ssssNNN...NNNRSSSSnnnnnnnnnnnnnnnnnnnnnnnn
```

Productive cDNA fragments for further analysis will contain adaptors
ligated at both ends.

84

5b. OPTIONAL: The ligated secondary index samples are incubated in the presence of phage T4 endonuclease VII, under mismatch cleavage conditions (Youil et al., 1996). This step will cleave any ligated DNA where the adaptors had been joined to an imperfectly paired sticky end. Thus, this step

5    performs error-checking of the previous ligation steps.

6. Each secondary index sample is separated into D aliquots, and D-1 of these aliquots (restricted secondary index samples) are digested with a four-base recognition restriction endonuclease that is not included in the set of the four-base recognition restriction endonucleases used in the first digest and the

10   mixture of N four-base recognition restriction endonucleases used in the second digest. The remaining undigested aliquot is the non-restricted secondary index sample. This digestion step will cleave a subset of the fragments and make them not amplifiable in step 7.

6b. (OPTIONAL): The material from each of the C * R * D secondary

15   index samples is placed on a separate microtiter well containing immobilized avidin (or any suitable binding moiety), in order to bind the biotinylated adapters (or any ligand-adapter) on the surface. This step will result in the immobilization of all the fragments that were successfully ligated to second adaptors in step 5, and were not cleaved in step 6. Any unbound fragments are

20   removed by washing. The resulting collection of C * R * D immobilized fragment sets is called the immobilized secondary index samples.

7. All of the immobilized secondary index samples are amplified using suitable amplification method, such as PCR. A total of C * R * D separate amplification reactions are performed. Two amplification primers are used, one

25   designed to be complementary to a strand of the second adaptor, and another designed to be complementary to the adaptor-indexer. The primer complementary to the adaptor-indexer is one of C possible primer sequences, overlapping by 4 bases with the enzyme recognition site of the first digest. The primer complementary to the second adaptor in each of the different R reactions

30   is one of N possible primer sequences, overlapping by 4 bases with each of the enzyme recognition sites of the mixture of N enzymes used in the second digest. Amplification is performed for a total of 12 to 25 cycles using PCR.

85

Example: Amplicons generated by PCR

```
second.adaptor-ssssNNN...NNNRSSSS-adaptor-indexer-3'
||||||||||||||||||||||||||||||||||||||||||||||||
second.adaptor-ssssNNN...NNNRSSSS-adaptor-indexer-5'
```

5        7b. There exist alternative methods for generating amplified tags in step

7. One example is the Strand Displacement Amplification method of Walker et

al., (1992). Yet another method for amplification of cDNA tags is rolling circle

amplification (Lizardi et al., 1998). In order to use rolling circle amplification, it

is necessary to generate circularized cDNA tags by ligation. Ligation methods

10     for generating circularized double stranded DNA are well known.

8. A digestion step with T7 exonuclease gene 6 (Nikiforov et al., 1993)

may be employed to generate single-stranded amplicons. This digestion step

assumes that one of the primers used in step 7 had been synthesized with 5

phosphorothioate linkages substituting for phosphodiester linkages at the 5' end

15     of the primer (Nikiforov et al., 1993). These linkages protect one strand from

digestion.

Example: Single-stranded amplicon after exonuclease digestion

```
second.adaptor-ssssNNN...NNNRSSSS-adaptor-indexer-5'
```

20       9. Each of the amplified secondary index samples is transferred to

another microtiter dish and hybridized with a fluorescence-labeled ligator-

detector oligonucleotide. There are C different ligator-detectors, each

complementary to one of the adaptor-indexers. Only the specific cognate

sequence is used in each of the amplified secondary index samples. The ligator-

25     detectors contain a 5'-phosphate group.

Example: cDNA amplicon hybridized with phosphorylated ligator-

detector

```
                       5'-RSSSS-ligator-detector-SIGNAL.A
                       ||||||||||||||||||||||||||
30   second.adaptor-ssssNNN...NNNRSSSS-adaptor-indexer-5'
```

9b. For more accurate expression profiling of control vs. tester mRNA

populations, this step may incorporate a second set of ligator-detectors labeled

with a fluorescent dye of a different color. Thus, control tags are hybridized

86

with C ligator-detectors of "signal type A", while tester tags are hybridized with C ligator-detectors of "signal type B". After hybridization is complete, paired sets of A-coded and B-coded secondary index samples are mixed, and one proceeds to step 10 below.

5          Example: cDNA amplicon hybridized with phosphorylated ligator-detector

```
                                  5'-RSSSS-ligator-detector-SIGNAL.B
                                   |||||||||||||||||||||||||
              second.adaptor-ssssNNN...NNNRSSSS-adaptor-indexer-5'
```

10

10. Each of the C * R * D secondary index samples, which now contain hybridized ligator-detectors, is contacted with a comprehensive address array (probe array). The comprehensive address array consists of a complete set of 4,096 hexamers (or, alternatively, 16,384 heptamers), ordered on a solid surface, and separated from the surface by a non-polynucleotide linker-spacer.

15  The hexamers (heptamers) contain free 3'-hydroxyl ends. Each of the C * R * D detection reactions on the arrays is performed under identical hybridization/ligation conditions, using T4 DNA ligase or any other suitable ligase, in order to achieve ligation of a specific, cognate hexamer sequence to

20  the corresponding sequence adjacent to the 5'-end of a ligator-detector probe. After hybridization, any unbound material is removed by very stringent washing.

Example: Ligator-detector joined to immobilized hexamer

Hexamer

```
25  SURFACE-linker-spacer-NNNNNN-3'

                                  5'-RSSSS-ligator-detector-
    SIGNAL.A
                                   |||||||||||||||||||||||||
30          second.adaptor-ssssNNN...NNNNNNRSSSS-adaptor-indexer-5'
```

Hexamer

```
    SURFACE-linker-spacer-NNNNNNRSSSS-ligator-detector-SIGNAL.A
                             |||||||||||||||||||||||||||||||||
35          second.adap-ssssNNN...NNNNNNRSSSS-adaptor-indexer-5'
```

87

Hexamer

```
SURFACE-linker-spacer-NNNNNNRSSSS-ligator-detector-SIGNAL.B
                      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
       second.adap-ssssNNN...NNNNNNRSSSS-adaptor-indexer-5'
```

5

11. The signals in each of the C * R * D comprehensive address arrays
are measured using a suitable readout instrument. An expression profile dataset
is generated from the signal ratios at each address of the C * R * D arrays. The
address of each signal is translated into a sequence, shown in bold letters. Six

10      nucleotides from the hexamer probe, five nucleotides from the
recognition/sticky end sequence, and one nucleotide from the last base in the
recognition sequence.

Example: Determined sequence for each array address.

                             6  +  5  + 1

15      second.adap-ssssNNN...NNNNNNRSSSS-adaptor-indexer-5'

Determined sequence:

            NNNNNNRSSSSR (10 nucleotides)


Specific signals that disappear in each one of the D-1 restricted

20      secondary index samples derived from a given secondary index sample are
indicative of the presence of a specific cleavage site [REST, below] within the
cDNA fragment.

                    4              4           6  +  5  + 1

       second.adap-ssssNNN...NNN[REST]NNN...NNNNNNRSSSS-adaptor-indexer

25          Determined sequence:

            ssss...[REST]...NNNNNNRSSSSR (4...4...10 = 18 nucleotides)

The expression dataset is analyzed taking into consideration the
expected address signatures of abundant gene products, which can be predicted
from available sequence data. The expression differences between two tissues

30      are measured by the ratio of the two different color signals corresponding to
control and tester. When N = 1, the sequence ssss sticky end sequence adjacent
to the second adaptor is also known.

88

## Examples

### Example 1 - Creation of FAAST-indexed DNA Amplicons

Fixed Address Analysis of Sequence Tags or FAAST sorts complex DNA into a collection of small fragments of partially known sequence by a process involving two sequential rounds of digestion with a restriction enzyme followed by ligation to duplex adapters. In the first round, DNA is digested with a II-S type restriction enzyme that cuts at a site distinct from the enzyme recognition site and generates a 5' or 3' overhang of undetermined sequence. Only duplex adapters that contain the complementary overhang will be capable of ligating to the restricted-DNA ends. In the second round, DNA fragments are digested with a restriction enzyme that cuts inside the recognition site and generates a 5' or 3' overhang of known sequence. Duplex adapters of corresponding sequence are subsequently ligated to these ends. In this example, for the first round of digestion we use FokI as the II-S enzyme, which cuts 9 bases from the recognition site and generates a 4-base 5' overhang. For the second round of digestion either NlaIII or DpnII can be used, which are 4-base restriction enzymes that cut within their corresponding recognition sites and generate 3' and 5' 4-base overhangs, respectively. Two yORFs of known sequence are subjected to the two rounds of digestion/ligation and demonstrate that the single PCR amplicon obtained from each yORF corresponds to the expected FAAST-indexed fragment.

In one procedure, two PCR products each containing a unique yeast open reading frame (yORF) were independently processed through the first phase of the FAAST procedure. This phase involves two sequential rounds of digestion with a restriction enzyme followed by ligation of the 4-base overhang produced by the enzymatic cleavage to a compatible adapter oligonucleotide complex. In the first round, carried out in solution, the yORFs were digested with FokI and the restriction ends were ligated to compatible adapter-indexer complexes. In the second round, carried out on a solid surface, the fragments were further digested with either NlaIII or DpnII and ligated to compatible second-adapter complexes. Two end products, one from each yORF, were

89

into solution. The elution volume (40 μl) was neutralized with 4 μl of 3 M

sodium acetate. Single-stranded DNA (ssDNA) was ethanol-precipitated using

5 μg of glycogen as carrier and redissolved in 10 μl of TE buffer. Eluted

ssDNA was not detectable using DNA-binding dyes in standard electrophoretic

5      analysis. 1 μl of each eluant was added as template into 50 μl PCR reactions

that included a primer pair complementary to indexer and second adapter

sequences. The FAAST-indexed amplicons expected from YCR100C and

YKL051W are 166 bp and 191 bp, respectively. DNA of the expected size was

apparent among the products of the corresponding PCR reaction. Samples from

10     PCR reactions started with dilutions of ssDNA as indicated above the lanes

were run on a 4-20% polyacrylamide gel in 1XTBE. Standards were used to

determine the size of indicated PCR products. Primer bands and putative PCR

artifacts were also identified. Digestion with restriction enzymes that recognize

a unique sequence within the predicted FAAST amplicons cleaved these

15     products into fragments of the expected size. DNA from PCR reactions started

with 1:10 dilutions of ssDNA were digested with the indicated restriction

enzyme. Bands corresponding to intact and restriction fragments were

identified. The 166 bp amplicon from YCR100C was cleaved by HpaII into 97

and 69 bp fragments. The 191 bp amplicon from YKL051W was cleaved by

20     HinfI into 111 and 80 bp fragments. The intact products were extracted from

the gel and used for further PCR amplification. Sequencing of this material

confirmed that it was composed of the expected FAAST amplicon sequences.

PCR primer annealing regions and 4-base restriction sites were identified. FokI

cleavage and recognition sites in that order were also identified. The same pair

25     of primers used for PCR amplification was used to sequence both strands of the

DNA fragments.

In summary, FAAST adapters were successfully ligated to specific

restriction fragments from two yORFs, as demonstrated by the production of the

expected PCR amplicons. These results demonstrate FAAST-indexing of

30     genomic DNA for use in subsequent hybridization analysis.

91

## CLAIMS

We claim:

1. A method of identifying nucleic acid fragments in a nucleic acid sample, the method comprising

(a) incubating a nucleic acid sample with one or more nucleic acid cleaving reagents that collectively generate sticky ends having a plurality of different sequences to produce nucleic acid fragments with sticky ends,

(b) mixing a plurality of adaptor-indexers with the nucleic acid sample and covalently coupling the adaptor-indexers to the nucleic acid fragments, wherein each adaptor-indexer has a different sticky end, wherein each sticky end of the adaptor-indexers is compatible with a sticky end generated by the nucleic acid cleaving reagents,

(c) hybridizing a plurality of ligator-detectors with the nucleic acid sample, wherein each ligator-detector comprises sequence matching or complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers,

(d) hybridizing a plurality of detector probes with the nucleic acid sample and covalently coupling the ligator-detectors to the detector probes, wherein each detector probe has a different sequence, and

(e) detecting, directly or indirectly, coupling of ligator-detectors to the detector probes.

2. The method of claim 1 wherein each adaptor-indexer comprises a sticky end portion and a detection portion, wherein the detector portion of each adaptor-indexer is separately detectable,

wherein the method further comprises detecting, directly or indirectly, detector portions of the adaptor-indexers.

3. The method of claim 2 wherein each ligator-detector comprises a detector portion, wherein the detector portion of each ligator-detector corresponds to the detector portion of one of the adaptor-indexers, wherein detection of the detection portion of a ligator-detector is an indirect detection of the detector portion of the corresponding adaptor-indexer.

93

the $c_0t_{1/2}$ for abundant nucleic acid fragments and less than the $c_0t_{1/2}$ for rare nucleic acid fragments, and collecting the un-renatured nucleic acid fragments.

14. The method of claim 8 wherein ligation of the ligator-detector is detected by rolling circle replication of an amplification target circle wherein replication is primed by the ligator-detector.

15. The method of claim 8 wherein the nucleic acid cleaving reagents generate sticky ends having N different sequences, and wherein the sample is divided into N index samples.

16. The method of claim 8 wherein the detector probes are all of the same length.

17. The method of claim 16 wherein the detector probes are six, seven, or eight nucleotides long.

18. The method of claim 8 wherein the detector probes all have similar hybrid stability.

19. The method of claim 8 wherein the nucleic acid cleaving reagents are restriction enzymes.

20. The method of claim 19 wherein the nucleic acid sample is digested with one restriction enzyme, wherein the restriction enzyme generates sticky ends having a plurality of different sequences.

21. The method of claim 20 wherein the restriction enzyme is a type II-S restriction enzyme that cleaves at a site different from its recognition sequence.

22. The method of claim 6 wherein each ligator-detector comprises sequence matching all or part of sequence including, and adjacent to, the sticky end of the adaptor-indexer.

23. The method of claim 6 wherein each ligator-detector comprises sequence complementary to all or part of sequence including, and adjacent to, the sticky end of the adaptor-indexer.

24. The method of claim 6 wherein at least one ligator-detector comprises sequence matching all or part of sequence including, and adjacent to, the sticky end of the adaptor-indexer and at least one ligator-detector comprises sequence complementary to all or part of sequence including, and adjacent to, the sticky end of the adaptor-indexer.

95

38. The method of claim 35 wherein the adaptor-indexers, ligator-detectors, or detector probes are sorted via the sorting tags.

39. The method of claim 35 wherein the ligator-detectors are associated with sorting tags, wherein the adaptor-indexers are associated with capture tags.

40. The method of claim 35 wherein the ligator-detectors are associated with capture tags, wherein the adaptor-indexers are associated with sorting tags.

41. The method of claim 35 wherein the detector probes are associated with sorting tags, wherein the adaptor-indexers are associated with capture tags.

42. The method of claim 35 wherein the adaptor-indexers, ligator-detectors, or detector probes are associated with a plurality of different capture tags or a plurality of different sorting tags.

43. The method of claim 8 further comprising determining the sequence of a portion of at least one of the nucleic acid fragments in the nucleic acid sample.

44. The method of claim 43 wherein the nucleic acid cleaving reagents are restriction enzymes,

wherein the nucleic acid sample is digested with one restriction enzyme,

wherein the restriction enzyme generates sticky ends having a plurality of different sequences,

wherein the restriction enzyme is a type II-S restriction enzyme that cleaves at a site different from its recognition sequence, and

wherein the portion of the nucleic acid fragments corresponds to the sticky end sequence, the sequence adjacent to the sticky end sequence to which the detector probe hybridized, and the recognition sequence of the restriction enzyme.

45. The method of claim 44 wherein the portion includes a gap of known length but unknown sequence between the sequence adjacent to the sticky end and the recognition sequence of the restriction enzyme.

46. The method of claim 45 wherein the portion has the structure

A-B-C-D

wherein A is the recognition sequence of the restriction enzyme, B is the gap of unknown sequence, C is the sequence to which the detector probe hybridized, and D is the sticky end sequence.

97

53. The method of claim 51 further comprising, prior to incubation with the second nucleic acid cleaving reagents,

dividing each index sample into a set of two or more of secondary index samples,

wherein each secondary index sample in each set of secondary index samples is incubated with a different set of one or more second nucleic acid cleaving reagents.

54. The method of claim 53 wherein steps (c), (d), and (e) are performed with each secondary index sample.

55. The method of claim 53 further comprising, prior to step (c),

dividing each secondary index sample into one or more restricted index samples and a non-restricted index sample,

incubating each restricted index sample with a different third nucleic acid cleaving reagent,

wherein steps (c), (d), and (e) are performed with both the restricted and non-restricted index samples.

56. The method of claim 55 wherein the first, second, and third nucleic acid cleaving reagents are restriction enzymes.

57. The method of claim 51 further comprising, following incubation with the second nucleic acid cleaving reagents,

dividing each index sample into a set of two or more of secondary index samples.

58. The method of claim 57 wherein steps (c), (d), and (e) are performed with each secondary index sample.

59. The method of claim 51 wherein the first and second nucleic acid cleaving reagents are restriction enzymes.

60. The method of claim 51 further comprising, following ligation of second adaptors to the nucleic acid fragments,

amplifying the nucleic acid fragments in the index samples to which adaptor-indexers and second adaptors have been ligated.

61. The method of claim 60 wherein the nucleic acid fragments are amplified by PCR.

99

71. The method of claim 8 further comprising, prior to ligation in steps (b) and (d),

mixing the index samples with a reagent that binds to base mismatches.

72. The method of claim 71 wherein the base mismatch reagent is (1) Mec1, (2) MutS, (3) MSH2-MSH6, (4) MLH1-PMS1, or (5) MSH2-MSH3.

73. The method of claim 8 wherein each adaptor-indexer, ligator-detector, or detector probe contains a label, wherein ligation of the ligator-detectors to the detector probes is detected via the label.

74. The method of claim 73 wherein each ligator-detector contains a label, wherein detecting ligation of the ligator-detectors to the detector probes is accomplished by

separating ligated ligator-detectors from unligated ligator-detectors, and

detecting the labels of the ligated ligator-detectors.

75. The method of claim 74 wherein each different ligator-detector contains a different label, wherein each detector probe is associated with a capture tag or a sorting tag, wherein separating ligated ligator-detectors from unligated ligator-detectors is accomplished by separating the detector probes from the unligated ligator-detectors using the capture tags or sorting tags, wherein the ligated ligator-detectors separate with the detector probes.

76. The method of claim 75 wherein the sorting tag is a fluorescent label, and wherein separating the detector probes from the unligated ligator-detectors is accomplished using a fluorescent label sorter.

77. The method of claim 73 wherein the labels are fluorescent, phosphorescent, or chemiluminescent labels.

78. The method of claim 77 wherein at least two of the labels are distinguished temporally via different fluorescent, phosphorescent, or chemiluminescent emission lifetimes.

79. The method of claim 73 wherein the labels are detectable by nuclear magnetic resonance, electron paramagnetic resonance, surface enhanced raman scattering, surface plasmon resonance, fluorescence, phosphorescence, chemiluminescence, resonance raman, microwave, or a combination.

80. The method of claim 79 wherein the label is detected using nuclear magnetic resonance, electron paramagnetic resonance, surface enhanced raman

threshold amount in the first nucleic acid sample but not present at the threshold amount in the second nucleic acid sample.

92. The method of claim 90 wherein the second nucleic acid sample is a sample from the same type of organism as the first nucleic acid sample.

93. The method of claim 90 wherein the second nucleic acid sample is a sample from the same type of tissue as the first nucleic acid sample.

94. The method of claim 90 wherein the second nucleic acid sample is a sample from the same organism as the first nucleic acid sample.

95. The method of claim 94 wherein the second nucleic acid sample is obtained at a different time than the first nucleic acid sample.

96. The method of claim 90 wherein the second nucleic acid sample is a sample from a different organism than the first nucleic acid sample.

97. The method of claim 90 wherein the second nucleic acid sample is a sample from a different type of tissue than the first nucleic acid sample.

98. The method of claim 90 wherein the second nucleic acid sample is a sample from a different species of organism than the first nucleic acid sample.

99. The method of claim 90 wherein the second nucleic acid sample is a sample from a different strain of organism than the first nucleic acid sample.

100. The method of claim 90 wherein the second nucleic acid sample is a sample from a different cellular compartment than the first nucleic acid sample.

101. The method of claim 90 further comprising identifying or preparing nucleic acid fragments corresponding the nucleic acid fragments present in the first nucleic acid sample but not present in the second nucleic acid sample.

102. The method of claim 101 further comprising using the nucleic acid fragments as probes.

103. The method of claim 102 wherein using the nucleic acid fragments as probes is accomplished by repeating steps (a) through (e) with a different nucleic acid sample, wherein the nucleic acid fragments are used as detector probes in steps (d) and (e).

detection of one or both strands of the ligated ligator-detector by mass spectroscopy.

112. The method of claim 106 further comprising, following ligation of the ligator-detectors to the detector probes,

incubation of the ligated ligator-detectors and detector probes with one or more nucleic acid cleaving reagents, and

detection of one or both strands of the ligated ligator-detector by mass spectroscopy.

113. The method of claim 8 further comprising, following step (b),

incubating the index samples with one or more second nucleic acid cleaving reagents, and

mixing a second adaptor with each index sample and ligating the second adaptors to the nucleic acid fragments, wherein each second adaptor has an end compatible with the end generated by one of the second nucleic acid cleaving reagents,

wherein at least one of the second nucleic acid cleaving reagents is sensitive to modification of its recognition site.

114. The method of claim 113 wherein the first and second nucleic acid cleaving reagents are restriction enzymes.

115. The method of claim 114 wherein the modification to the recognition site is methylation, alkylation, dimerization, derivatization, depurination, or ADP-ribosylation.

116. The method of claim 114 wherein the modification is present in the nucleic acid fragments when isolated or is introduced to the nucleic acid fragments after isolation.

117. The method of claim 114 further comprising, prior to step (c),

dividing each index sample into one or more restricted index samples and a non-restricted index sample,

digesting each restricted index sample with a different third restriction enzyme,

wherein steps (c), (d), and (e) are performed with both the restricted and non-restricted index samples.

105

123. The method of claim 114 wherein the pattern of the presence, amount, presence and amount, or absence of ligator-detectors ligated to different detector probes constitutes a catalog of nucleic acid fragments in the nucleic acid sample.

124. The method of claim 123 further comprising preparing a second catalog of nucleic acid fragments in a second nucleic acid sample and comparing the first catalog and second catalog, wherein differences in the first and second catalogs indicate differences in modification of the first and second nucleic acid samples.

125. The method of claim 123 wherein the second nucleic acid sample is a sample from the same type of cells as the first nucleic acid sample except that the cells from which the first nucleic acid sample is derived are modification-deficient relative to the cells from which the second nucleic acid sample is derived.

126. The method of claim 123 wherein the second nucleic acid sample is a sample from a different type of cells than the first nucleic acid sample, and wherein the cells from which the first nucleic acid sample is derived are modification-deficient relative to the cells from which the second nucleic acid sample is derived.

127. The method of claim 6 wherein the nucleic acid cleaving reagents are restriction enzymes, wherein the restriction enzymes are type II-S restriction enzymes that cleave at sites different from their recognition sequences,

the method further comprising, prior to step (a),

introducing a recognition site for the restriction enzyme adjacent to a region of interest in one or more nucleic acid molecules in the nucleic acid sample,

wherein the ligator-detector comprises sequence matching or complementary to all or part of the sequence of, and adjacent to, the recognition site for the restriction enzyme.

128. The method of claim 127 wherein the recognition site is introduced by amplifying the nucleic acid molecules using a primer containing the recognition sequence.

129. The method of claim 128 wherein the nucleic acid molecules are amplified using PCR.

130. The method of claim 127 wherein the detector probes include sequence matching or complementary to known or potential sequence in the region of interest.

131. The method of claim 130 wherein there are a plurality of known or potential sequences in the region of interest, wherein the set of detector probes

107

138. The kit of claim 137 wherein at least two of the labels are distinguished temporally via different fluorescent, phosphorescent, or chemiluminescent emission lifetimes.

139. The kit of claim 136 wherein the labels are detectable by nuclear magnetic resonance, electron paramagnetic resonance, surface enhanced raman scattering, surface plasmon resonance, fluorescence, phosphorescence, chemiluminescence, resonance raman, microwave, or a combination.

140. The kit of claim 136 wherein the labels are beads comprising a label.

141. The method of claim 140 wherein the label is a molecular barcode.

142. The kit of claim 136 wherein the labels are mass labels.

143. The kit of claim 133 wherein the nucleic acid cleaving reagent generates sticky ends having N different sequences, and wherein the kit includes N different adaptor-indexers.

144. The kit of claim 133 wherein the detector probes are six, seven, or eight nucleotides long.

145. The kit of claim 133 wherein each ligator-detector contains a label.

146. The kit of claim 145 wherein the label is a fluorescent label.

147. The kit of claim 145 wherein the kit includes at least two ligator-detectors corresponding to each adaptor-indexer, wherein each ligator-detector corresponding to a given adaptor-indexer contains a different label.

148. The kit of claim 133 wherein the detector probes are all of the same length.

149. The kit of claim 133 wherein the detector probes all have similar hybrid stability.

150. The kit of claim 133 wherein each adaptor-indexer, ligator-detector, or detector probe is immobilized on a substrate.

151. The kit of claim 150 wherein the detector probes are immobilized on a substrate.

152. The kit of claim 133 wherein each adaptor-indexer, ligator-detector, or detector probe is associated with a capture tag, sorting tag, or both.

153. The kit of claim 152 wherein the ligator-detectors are associated with capture tags or sorting tags.

109

## Figure 1

| Adaptor-indexer I | Adaptor-indexer II |
|---|---|
| CATGCGGATCCTAAGGCTTACGCC | TAATGGAAGCTGGATTCGCG |
| GCCTAGGATTCCGAATGCGG | GTTCATTACCTTCGACCTAAGCGC |

| Ligator-detectors | Ligator-detectors |
|---|---|
| 1.  CATGCGGATCCTAAGGCTTACGCC | 21.  CAAGTAATGGAAGCTGGATTCGCG |
| 2.  CATGCGGATCCTAAGGCTTACGC | 22.  CAAGTAATGGAAGCTGGATTCGC |
| 3.  CATGCGGATCCTAAGGCTTA | 23.  CAAGTAATGGAAGCTGGATTC |
| 4.  CATGCGGATCCTAAGGC | 24.  CAAGTAATGGAAGCT |
| 5.  ATGCGGATCCTAAGGCTTACGCC | 25.  AAGTAATGGAAGCTGGATTCGCG |
| 6.  TGCGGATCCTAAGGCTTACGCC | 26.  AGTAATGGAAGCTGGATTCGCG |
| 7.  GCGGATCCTAAGGCTTACGCC | 27.  GTAATGGAAGCTGGATTCGCG |
| 8.  ATGCGGATCCTAAGGCTTACGC | 28.  AAGTAATGGAAGCTGGATTCGC |
| 9.  ATGCGGATCCTAAGGCTTACG | 29.  AAGTAATGGAAGCTGGATTC |
| 10.  ATGCGGATCCTAAGGCTT | 30.  AAGTAATGGAAGCTGGAT |
| 11.  ATGCGGATCCTAAGGC | 31.  AAGTAATGGAAGCTG |
| 12.  TGCGGATCCTAAGGCTTAC | 32.  AGTAATGGAAGCTGGATTCGCG |
| 13.  GGCGTAAGCCTTAGGATCCGCATC | 33.  CGCGAATCCAGCTTCCATTACTTG |
| 14.  GCGTAAGCCTTAGGATCCGCATC | 34.  GCGAATCCAGCTTCCATTACTTG |
| 15.  GTAAGCCTTAGGATCCGCATC | 35.  GAATCCAGCTTCCATTACTTG |
| 16.  CCTTAGGATCCGCATC | 36.  CAGCTTCCATTACTTG |
| 17.  GGCGTAAGCCTTAGGATCCGCAT | 37.  CGCGAATCCAGCTTCCATTACTT |
| 18.  GGCGTAAGCCTTAGGATCCGCA | 38.  CGCGAATCCAGCTTCCATTACT |
| 19.  GGCGTAAGCCTTAGGATCCGC | 39.  CGCGAATCCAGCTTCCATTAC |
| 20.  GCATGCGGATCCTAAGGCTTACGCC | 40.  GAATCCAGCTTCCATTACTT |

Figure 3

A  ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGACAACCAGCCATTCACGGGC...
     ...TATACGAAATCCGCCTACGATTTAGCATACCTGTTGGTCGGTAAGTGCCCG...

B  ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGT     ATGGACAACCAGCCATTCACGGGC...
     ...TATACGAAATCCGCCTACGATTTAGCATACC      TGTTGGTCGGTAAGTGCCCG...

C  ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGT     ATGGCGGATCCTAAGGCTTACGCC
     ...TATACGAAATCCGCCTACGATTTAGCATACC      GCCTAGGATTCCGAATGCGG

D  ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC
     ...TATACGAAATCCGCCTACGATTTAGCATACCGCCTAGGATTCCGAATGCGG

E                     ATGGCGGATCCTAAGGCTTACGCC-label

     ...TATACGAAATCCG<u>CCTAC</u>GATTTAGCATACCGCCTAGGATTCCGAATGCGG

F.                   ATGGCGGATCCTAAGGCTTACGCC-label
     ...TATACGAAATCCG<u>CCTAC</u>GATTTAGCATACCGCCTAGGATTCCGAATGCGG

G          support
               AATCGT

                        ATGGCGGATCCTAAGGCTTACGCC-label
   ...TATACGAAATCCG<u>CCTAC</u>GATTTAGCATACCGCCTAGGATTCCGAATGCGG

H          support
               AATCGTATGGCGGATCCTAAGGCTTACGCC-label
   ...TATACGAAATCCG<u>CCTAC</u>GATTTAGCATACCGCCTAGGATTCCGAATGCGG

I          support
               AATCGTATGGCGGATCCTAAGGCTTACGCC-label

Figure 5

A    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGACAACCAGCCATTCACGGGC...
      ...TATACGAAATCCGCCTACGATTTAGCATACCTGTTGGTCGGTAAGTGCCCG...


B    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGT      ATGGACAACCAGCCATTCACGGGC...
      ...TATACGAAATCCGCCTACGATTTAGCATACC      TGTTGGTCGGTAAGTGCCCG...


C    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGT      ATGGCGGATCCTAAGGCTTACGCC-tag
      ...TATACGAAATCCGCCTACGATTTAGCATACC      GCCTAGGATTCCGAATGCGG


D    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC-tag
      ...TATACGAAATCCGCCTACGATTTAGCATACCGCCTAGGATTCCGAATGCGG


E    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC-tag

                               TACCGCCTAGGATTCCGAATGCGG-label


F.    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC-tag
                               TACCGCCTAGGATTCCGAATGCGG-label


G    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC-tag
                               TACCGCCTAGGATTCCGAATGCGG-label

                  TTAGCA
       support


H    ...ATATGCTTTAGGC<u>GGATG</u>CTAAATCGTATGGCGGATCCTAAGGCTTACGCC-tag
                  TTAGCATACCGCCTAGGATTCCGAATGCGG-label
       support


I                   TTAGCATACCGCCTAGGATTCCGAATGCGG-label
       support

(54) Title: FIXED ADDRESS ANALYSIS OF SEQUENCE TAGS

(57) Abstract: Disclosed is a method for the comprehensive analysis of nucleic acid samples and a detector composition for use in the method. The method, referred to as Fixed Address Analysis of Sequence Tags (FAAST), involves generation of a set of nucleic acid fragments having a variety of sticky end sequences; indexing of the fragments into sets based on the sequence of sticky ends; associating a detector sequence with the fragments; sequence-based capture of the indexed fragments on a detector array; and detection of the fragment labels. Generation of the multiple sticky end sequences is accomplished by incubating the nucleic acid sample with one or more nucleic acid cleaving reagents. The indexed fragments are captured by hybridization and coupling, preferably by ligation, to a probe. The method allows a complex sample of nucleic acid to be quickly and easily cataloged in a reproducible and sequence-specific manner. One form of the method allows determination of associations, in a nucleic acid molecule, of different combinations of known or potential sequences. Another form of the method assesses modification of sequences in nucleic acid molecules by basing cleavage of the molecules on the presence or absence of modification.

# INTERNATIONAL SEARCH REPORT

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | WO 97 46704 A (LYNX THERAPEUTICS INC) 11 December 1997 (1997-12-11) claims 1-27 | 1-154 |
| Y | WO 97 13877 A (LYNX THERAPEUTICS INC ;MARTIN DAVID W (US)) 17 April 1997 (1997-04-17) page 23, line 10 -page 33, line 10 | 1-154 |
| Y | WO 98 30721 A (PIONEER HI BRED INT ;BIRO RONALD L (US); FEAZEL RHONDA (US); HELEN) 16 July 1998 (1998-07-16) claims 1,36; examples 1-3 | 1-154 |
| Y | WO 93 17126 A (NEW YORK HEALTH RES INST) 2 September 1993 (1993-09-02) claims 1,36; examples 1-3 | 1-154 |
| Y | WO 98 31831 A (BRAX GENOMICS LTD ;THOMPSON ANDREW HUGIN (GB); SCHMIDT GUNTER (GB)) 23 July 1998 (1998-07-23) page 8 -page 16; claim 1 | 1-154 |
| Y | WO 97 31256 A (BLOK HERMAN ;BARANY GEORGE (US); KEMPE MARIA (US); ZIRVI MONIB (US) 28 August 1997 (1997-08-28) claims 1,2 | 1-154 |
| A | WO 98 48047 A (BRAX GENOMICS LTD ;THOMPSON ANDREW HUGIN (GB); SCHMIDT GUENTER (GB) 29 October 1998 (1998-10-29) the whole document | 1-154 |
| P,X | WO 00 09756 A (PERKIN ELMER CORP) 24 February 2000 (2000-02-24) claim 1 | 1-154 |
| E | WO 01 12855 A (UNIV YALE) 22 February 2001 (2001-02-22) the whole document | 1-154 |

2

| Patent document cited in search report | | Publication date | Patent family member(s) | | Publication date |
|---|---|---|---|---|---|
| WO 9713877 | A | | US 6280935 B1 | | 28-08-2001 |
| | | | WO 9641011 A1 | | 19-12-1996 |
| | | | WO 9713877 A1 | | 17-04-1997 |
| | | | US 6235475 B1 | | 22-05-2001 |
| | | | AU 718357 B2 | | 13-04-2000 |
| WO 9830721 | A | 16-07-1998 | AU 6024598 A | | 03-08-1998 |
| | | | US 6100030 A | | 08-08-2000 |
| | | | WO 9830721 A1 | | 16-07-1998 |
| WO 9317126 | A | 02-09-1993 | AU 3728093 A | | 13-09-1993 |
| | | | CA 2130562 A1 | | 02-09-1993 |
| | | | EP 0675966 A1 | | 11-10-1995 |
| | | | WO 9317126 A1 | | 02-09-1993 |
| | | | US 6103463 A | | 15-08-2000 |
| WO 9831831 | A | 23-07-1998 | AU 728805 B2 | | 18-01-2001 |
| | | | AU 5570298 A | | 07-08-1998 |
| | | | EP 0960215 A1 | | 01-12-1999 |
| | | | WO 9831831 A1 | | 23-07-1998 |
| | | | JP 2001508308 T | | 26-06-2001 |
| WO 9731256 | A | 28-08-1997 | AU 735440 B2 | | 05-07-2001 |
| | | | AU 2799797 A | | 10-09-1997 |
| | | | CA 2244891 A1 | | 28-08-1997 |
| | | | EP 0920440 A2 | | 09-06-1999 |
| | | | WO 9731256 A2 | | 28-08-1997 |
| WO 9848047 | A | 29-10-1998 | AU 733924 B2 | | 31-05-2001 |
| | | | AU 7063898 A | | 13-11-1998 |
| | | | CN 1252103 T | | 03-05-2000 |
| | | | EP 0981645 A1 | | 01-03-2000 |
| | | | WO 9848047 A2 | | 29-10-1998 |
| WO 0009756 | A | 24-02-2000 | US 6232067 B1 | | 15-05-2001 |
| | | | AU 5560199 A | | 06-03-2000 |
| | | | WO 0009756 A1 | | 24-02-2000 |
| | | | US 6258539 B1 | | 10-07-2001 |
| WO 0112855 | A | 22-02-2001 | US 6261782 B1 | | 17-07-2001 |
| | | | AU 6638000 A | | 13-03-2001 |
| | | | AU 6770800 A | | 13-03-2001 |
| | | | AU 6900100 A | | 13-03-2001 |
| | | | WO 0112671 A1 | | 22-02-2001 |
| | | | WO 0112855 A2 | | 22-02-2001 |
| | | | WO 0112856 A2 | | 22-02-2001 |

THIS PAGE BLANK (USPTO)